To the Graduate Council:

I am submitting herewith a dissertation written by Cheryl R. Hild entitled "Development of the Group Method of Data Handling with Information-Based Model Evaluation Criteria: A New Approach to Statistical Modeling." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Management Science.

Hamparsum Bozdogan,
Major Professor

We have read this dissertation
and recommend its acceptance:

Accepted for the Council:

Associate Vice Chancellor and
Dean of The Graduate School

DEVELOPMENT OF THE GROUP METHOD OF DATA HANDLING WITH INFORMATION-BASED MODEL EVALUATION CRITERIA: A NEW APPROACH TO STATISTICAL MODELING

A Dissertation
Presented for the
Ph.D. Degree in Management Science
The University of Tennessee, Knoxville

CHERYL R. HILD
August, 1998

"The only real progress lies in learning to be wrong all alone." -Albert Camus

I dedicate this thesis to my mother, Anna Marlene Hild. Without her inspiration, its progress would have never found an end. By example, she taught me persistence, dedication, and patience.

ACKNOWLEDGMENTS

There are many individuals who have had great influence on my career and educational choices. These persons have provided me with guidance, inspiration, and wisdom. To each of these, I am forever grateful.

First, I would like to acknowledge my teacher, friend, and thesis advisor, Dr. Hamparsum Bozdogan. In addition to introducing me to the topic of this thesis, he has consistently supported me throughout my research endeavors.

Secondly, I thank Dr. Mary Leitnaker, who has been a very special friend as well as a member of my committee. She helped to make many of the onerous times bearable.

I am thankful to Dr. Kenneth Gilbert for his positive influence on my decision to pursue my graduate studies at The University of Tennessee's College of Business.

I am extremely appreciative for Professor John Neter, for whom I have great admiration. My interests in the area of statistical modeling was sparked while I was studying under him at the University of Georgia.

Thanks are also due to my undergraduate advisor and professor, Dr. Joseph G. VanMatre. His passion for statistics influenced my decision regarding my undergraduate degree. If it had not been for him, I would have a B.S. in accounting. Thank you, Joe.

I also thank John Riblett of the Management Development Center not only for the financial support he provided during my graduate program, but also for the many opportunities for growth and development.

I thank my family members -- my father for his patience and support through my many years of school and my sister for her listening ear. I will forever be thankful to my brother for always believing in me. Most importantly, I must thank my husband, Bobby, and my precious son, Lee, for their sacrificial love and support.

ABSTRACT

A new approach to statistical modeling is studied which integrates the *information-based statistical criteria* into Ivakhnenko's (1966) polynomial theory, also known as the *Group Method of Data Handling* (GMDH).

The GMDH algorithm was introduced as a hierarchical algorithm which constructs a mathematical model of a complex system by the composition of lower-order polynomials of the form:

$$z = A + Bx + Cy + Dx^2 + Ey^2 + Fxy.$$

Under the assumptions that there exists a random error component additive with the above deterministic component, the polynomials generated at each level of the GMDH algorithm are statistically modeled as regression-type equations of the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon.$$

In this research, the assumption of independence of error terms is relaxed by the derivation of information-based model selection criteria for the GMDH algorithm. These criteria inherently guard against multicollinearity by incorporating a measure of the complexity of each model being evaluated. In addition to deriving the information-based criteria for the GMDH, the consistency of the criteria when used within the GMDH algorithm is established. While consistency is a desirable asymptotic property for model evaluation, it does not, by itself, guarantee that the criteria are effective in terms of model selection. Thus, Monte Carlo experimentation is used to study the aptness (i.e., the ability of the model to describe the system of interest) of the model identified by the GMDH algorithm using information-based criteria.

At the end of the GMDH algorithm, the "best approximating" model is identified. In generations two and beyond, the models are written in terms of the previous generation's polynomials. Hence, the chosen model is typically not written in terms of the original set of predictor variables. Even with the original GMDH algorithm, existing computer programs do not provide a means of backtracing

(i.e., back-substituting) to the original variables. The revised algorithm (using the information-based model evaluation criteria) is coded in MATLAB®. A backtracing algorithm is introduced and designed for the GMDH hierarchical tree-algorithm to determine the "best" fitting model in terms of the original variables using a symbolic mathematical language by borrowing on the rules of algebraic manipulation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND PRELIMINARIES

1.1 Introduction to Statistical Modeling

The development and use of mathematical models has been an important topic in research institutions and industry for centuries. These mathematical models, which include deterministic and statistical models, address the need to be able to identify the critical relationships between variables that impact the behavior of systems and to describe or predict the behaviors of such systems. For example, the Internal Revenue Service (IRS) is responsible for identifying potential non-compliances, conducting audits of those cases, and determining the proper tax adjustments on each audit. (U.S. General Accounting Office, 1996) Members of the research branch of the IRS are asked to determine factors, including taxpayer characteristics and IRS policies and procedures, on compliance. Can certain factors be identified which can be used to predict compliance

and yield levels? The objective is to be able to express the uncertainties of this type of problem in terms of a mathematical model. Obviously, no set of factors will provide a deterministic relationship between taxpayer and IRS procedural characteristics and the compliance or yield levels.

1.1.1 Definitions and Basic Concepts

Certain concepts that are foundational to the area of modeling and terms that define objects or phenomena under study are used throughout this thesis. These terms and concepts are now defined.

Definition: A *functional relation* between two variables is a relationship that can be expressed completely by a mathematical formula.

Based on this definition, if two variables, x and y, have a functional relation, if the value of one of the variables is known (say, x), then the value of y is also known. A functional relation between the variable x and y is typically expressed in the form $y=f(x)$, where

given a value of x, the function *f* indicates the corresponding value of y.

Definition: A *mathematical model* is a mathematical formulation that is used to represent or approximate the behavior or output(s) of a system.

Mathematical models can be extremely simple, with only a single variable, or highly complex, with many variables and representative equations. A mathematical model is based on the fact that there is a distinct functional relationship between the variables. Also, the term *mathematical model* encompasses both deterministic models and statistical models.

Definition: A *statistical model* is a mathematical formulation that expresses the uncertainties surrounding the functional relation in terms of probabilities.

A statistical model is used in attempts to approximate the random behavior of systems, since it is extremely rare for a single functional

relation to completely define the outputs of a system. Thus, a statistical model takes into account uncertainties in the relationship between the variables used to characterize a system's stochastic behavior.

Definition: A *parameter* of a model is a constant used to represent one of the characteristic variables of a system.

One of the major objectives in mathematical modeling is to determine the correct or "best" values of these model parameters. Once the parameter values are determined, then the value of a system output (i.e., a *dependent* or *response* variable) can be predicted given values for the variables used to characterize the system (i.e., *independent* or *predictor* variables).

1.1.2 Uses and Selection of Statistical Models

In contrast to the deterministic model, non-deterministic, statistical models take into account the deterministic as well as the random component of a system. Thus, a statistical model typically distinguishes between the functional and statistical relations

between the response variable and the independent or predictor variables. Statistical models serve two primary purposes--to provide a description of the system from which the data is collected and to allow for prediction of the behavior of the system. Examples of statistical models include regression analysis, time series models, and analysis of variance models.

The appropriate statistical modeling method is based on the type of information desired from the model. For instance, analysis of variance models are used in experimentation in order to determine how much of the variability in a system output is attributable to specific independent variables and how much of the variability is due to other unknown factors. On the other hand, regression analysis is concerned with determining the form of the relationship between a response variable and one or more independent variables so that the model can be used to predict future values of the response variable. Thus, regression analysis must address such questions as: What variables should be included in the model? Is the form of the relationship between the variables linear? If the relationship is not linear, how can the form be described in a mathematical equation? Finally, time series analysis is a set of modeling techniques which

attempt to develop a mathematical equation that models the behavior of a variable over time. Thus, times series models sometimes use regression analysis to estimate the trend of a variable over time, but they also attempt to account for seasonal and other patterns occurring over time. In general, modeling of time series data is not a trivial exercise.

There exist three main steps in the modeling of data (including times series data). These are: (i) model specification or identification, (ii) model fitting, and (iii) model diagnostics. Even though there are many different types and uses of statistical models, there is one main question that underlies the development of any statistical model:

> *Given a set of data from a system of interest, what is the appropriate distributional model to describe the statistical relation and what is the appropriate functional form of the model to adequately represent the underlying mechanism or system from which the data are obtained?*

1.2 The Predictive Model of Interest

Let y represent the response or dependent variable and $x=(x_1,x_2,...,x_m)$ be the set of m independent or predictor variables. Then, the general form of the predictive model of interest is:

$$y = f(x_1,x_2,...,x_m) + \varepsilon, \tag{1.1}$$

where $f(x_1,x_2,...,x_m)$ is an unknown mathematical function relating to the deterministic component of the model and ε is the random error relating to the statistical component of the model. The predictive model of equation (1.1) is a generalized form of the regression-type model where $f(x_1, x_2, ..., x_m)$ is commonly taken to be the linear combination of x_1, x_2, ..., x_m. This general linear model is typically represented in matrix form as:

$$\underset{(nx1)}{y} = \underset{(nxp)}{X}\ \underset{(px1)}{\beta} + \underset{(nx1)}{\varepsilon}, \tag{1.2}$$

where:

$p = m+1$, the number of parameters in the model,

y is a (nx1) vector of observations for the response variable,

X is a (nxp) matrix of values for the independent variables,

β is a (px1) vector of parameters, and

ε is the (nx1) vector of random errors.

Since model (1.2) is the matrix form of the regression model, a data structure is used like that of multiple regression analysis, as shown in Figure 1.1.

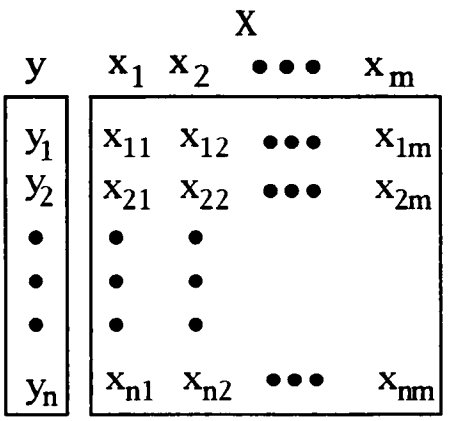

FIGURE 1.1. Data Structure for Predictive Model of Interest

1.3 Traditional Statistical Modeling Techniques

Statistical modeling includes the identification of an appropriate model for a given realization, estimation of the values of the parameters of the identified model, validation of the model, and the final application of the model. Traditionally, these components of statistical modeling are performed separately and sequentially, as Figure 1.2 illustrates. Akaike (1974) and Bozdogan (1981) argue that, in traditional statistical modeling methods, there is a meaningless separation of the estimation and validation elements. They present information-based statistical criteria as a means for integrating these components of statistical modeling. We propose a heuristical modeling technique combined with information-based criteria to integrate these components so that the identification of the model, the estimation of parameter values, and the validation of the model are treated as one entity. However, it is useful to first review some of the assumptions and techniques associated with traditional statistical modeling.

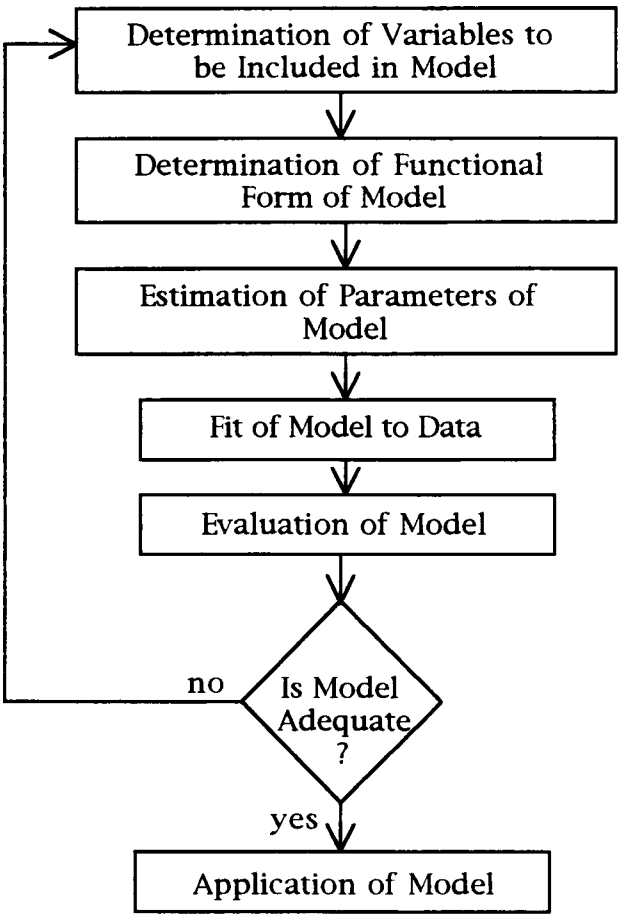

FIGURE 1.2. Traditional Statistical Modeling Process

1.3.1 Traditional Modeling Assumptions

The assumptions associated with many traditional statistical modeling techniques (e.g., regression analysis) are:

1) The functional form of the model is known prior to the modeling of the data.
2) The deterministic and random components of the model are additive.
3) The probability distributions of the response variable have the same variance, σ^2, regardless of the values of the predictor variables.
4) The error terms are uncorrelated. Hence, the responses y_i and y_j $(i \neq j)$ are uncorrelated.
5) The random error terms are normally distributed.

It is fairly obvious that several of these assumptions are impractical. For instance, prior to the collection and study of data on a system, it is highly unlikely that the functional form of the model is known. We propose an approach that allows for the relaxation of the first assumption and that permits various distributional assumptions for the error terms.

1.3.2 Definition of the Univariate Normal Distribution

Many statistical modeling techniques are based on an assumption that the system output of interest can be reasonably approximated by a normal probability distribution. This assumption is attractive since the normal distribution has unique mathematical properties. Within the GMDH framework, we will initially develop our approach based on this assumption of normality. For convenience, we now define the normal distributional model for our predictive model of interest defined in (1.2).

If X is a normally distributed random variable, then the probability density function for X is:

$$f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2 \right] \quad -\infty < X < \infty , \qquad (1.3)$$

where μ and σ are the parameters of the normal distribution and exp(a) denotes e^a. The mean and variance of the normal random variable X are:

$$E(X) = \mu \tag{1.4}$$

$$\sigma^2(X) = \sigma^2.$$

We denote that X is normally distributed with mean μ and variance σ^2 by $X \sim N(\mu, \sigma^2)$. One of the nice properties of a normally distributed random variable, X, is that a linear function of X is also normally distributed. Therefore, if X is a normal random variable and Y = a + bX, where a and b are constants, then Y is normally distributed with mean a+bE(X) and variance $b^2\sigma^2(X)$.

When the functional form of the probability distribution of the random error terms in (1.2) is specified to be normal, then the following expectation is defined for known (i.e., fixed) x's:

$$E(y_i) = \underset{(1xp)}{x_i'} \underset{(px1)}{\beta}, \qquad i=1,2,...,n \tag{1.5}$$

The normal error model is thus specified as:

$$f(y_i \mid x_i) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left(y_i - x_i'\beta\right)^2\right] \qquad (1.6)$$

The likelihood function for the normal model (1.6) is simply the joint probability function of the n sample observations:

$$L(\beta, \sigma^2) = \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2\right] \qquad (1.7)$$

The maximum likelihood estimators of the parameters are those which maximize the likelihood function (1.7). For the normal model, the maximum likelihood estimators, $b_0, b_1, \ldots, b_m$ of $\beta_0, \beta_1, \ldots, \beta_m$ are the same as those estimators provided by the method of least squares. The maximum likelihood estimators $b_0, b_1, \ldots, b_m$ are unbiased and consistent estimators and they have minimum variance among all unbiased linear estimators.

1.3.3 Estimation of Parameters

One of the components of statistical modeling is estimating the unknown values of the parameters of the model. *Estimation* uses

sample observations to obtain estimates of the true values of the parameters from the available data. There are two general methods that are commonly used in statistical modeling for estimating parameter values - the method of maximum likelihood and the method of least squares. Whatever the method of estimation, there is obviously some error involved due to random error and misspecification of the model. Several properties of estimators are traditionally used to determine how well the estimated values represent the true population parameters.

Let $x_1, x_2, ..., x_n$ represent a random sample of n observations and let $\theta = (\theta_1, \theta_2, ..., \theta_k)$ represent a set of k population parameters to be estimated.

Definition: An estimator $\hat{\theta}$ of the parameter θ is *unbiased* if $E(\hat{\theta}) = \theta$.

Definition: $\hat{\theta}$ is a *consistent* estimator of θ if $\lim_{n \to \infty} P(|\hat{\theta} - \theta| \geq \varepsilon) = 0$ for any $\varepsilon > 0$.

Definition: $\hat{\theta}$ is a *minimum variance* estimator of θ if, for any other estimator θ', $\sigma^2(\hat{\theta}) \leq \sigma^2(\theta')$.

As previously stated, there exist two general methods that are traditionally used to estimate the parameters of a model. The method of maximum likelihood is well established in introductory mathematical statistics books, such as Lindgren (1976). The maximum likelihood method is based on the *maximum likelihood principle*, which basically states that the estimator $\hat{\theta}$ that provides the most information from a set of data is the value that maximizes the likelihood function,

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta). \tag{1.8}$$

Under certain conditions of regularity, maximum likelihood estimators are consistent estimators and they are asymptotically normally distributed. (Lindgren, 1976)

The *method of least squares* is another general method of finding estimators when the sample observations are of the form $y_i = f(\theta_1,\ldots,\theta_k) + \varepsilon_i$ for $i = 1,\ldots,n$. As presented by Neter, Wasserman, and Kutner (1983), the least squares estimator of θ_j $(j=1,\ldots,p)$ is obtained by minimizing

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [(y_i - f(\theta_1,\ldots,\theta_p)]^2 \qquad (1.9)$$

with respect to θ_j. Thus, the method of least squares does not require that the probability distribution of the error terms be known. Least squares estimators usually have the property of being unbiased.

1.3.4 Traditional Model Evaluation Criteria

In addition to obtaining good parameter estimates, it is also necessary to evaluate the model as a whole in terms of how well the model approximates the behavior of the system of interest. In order to develop a "good", predictive statistical model, three main characteristics of the model must be considered:

1) The goodness-of-fit of the developed model to the actual system,
2) the variation that exists in the parameter estimates, and
3) the parsimoniousness and complexity of the model.

Here, parsimoniousness is based on the *principle of parsimony* which states that a complex model should not be used when a simpler model is sufficient. Empirically, *complexity* is defined to be a measure of the interdependence among the components of a random vector.

In most classical statistical modeling methods, such as regression analysis, model evaluation criteria are based on inferences about population means from normal distributions. Thus, assuming that the normal random error model is applicable, the significance of the parameter estimates and the goodness of fit of the model is typically based on the most commonly used measure of performance, the *mean squared error* (MSE):

$$\text{MSE} = \frac{\sum (y_i - \hat{y}_i)^2}{n - p}, \tag{1.10}$$

where y_i is the observed value of the response variable, $\hat{y}_i$ is the predicted value of y_i using the estimated model, and p is the number of parameters estimated in the model. Based on the mean squared

error, an F test is used to test whether there is a regression relation between the dependent variable, y, and the set of predictor variables, $x_1,...,x_m$.

Also related to the mean square error is the coefficient of multiple determination, R^2, which is defined as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}. \qquad (1.11)$$

R^2 is the most commonly used criterion for deciding on the appropriateness of the chosen model. However, a large R^2 does not necessarily imply that the estimated model is useful for prediction. Since R^2 increases with the addition of more predictor variables, many practitioners utilize the adjusted coefficient of determination, R_a^2, which takes into account both the sample size as well as the number of parameters in the model. R_a^2 is defined as:

$$R_a^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}. \qquad (1.12)$$

When multicollinearity exists among the independent variables, these traditional evaluation methods become weaker. *Multicollinearity*, or intercorrelation, exists when the independent variables are correlated among themselves. A conclusion concerning a regression relation based on an F test as well as the value of the coefficient of determination are affected by the existence of multicollinearity. Thus, the methodology presented in this thesis utilizes information-based criteria in addition to these traditional evaluation methods. Information-based model evaluation criteria provide a means for effectively evaluating the usefulness of a model even when multicollinearity exists among the independent variables. Chapter 3 discusses the theory and the practical application of the use of information-based criteria in model identification and evaluation.

1.4 An Example of Estimating the Predictive Model of Interest Using Regression Analysis

We consider the multiple regression analysis of data from a study performed by J. T. Wakely (1949) on the quantity of vitamin

B_2 in turnip greens. These data are provided in Table 1.1. Anderson and Bancroft (1959) fit the following predictive model to these data:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \varepsilon, \qquad (1.13)$$

where:

x_1 = radiation in gram calories per minute during preceding half day of sunlight,

x_2 = average soil moisture tension,

x_3 = air temperature in degrees Fahrenheit, and

y = milligrams of vitamin B_2 per gram of turnip green.

Regression analysis under model (1.13) yields the parameter estimates provided in Table 1.2 and the summary of fit in Table 1.3. However, if we fit the polynomial model,

$$y = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 x_2^2 + \varepsilon \qquad (1.14)$$

TABLE 1.1. Data for Study of Vitamin B_2 in Turnip Greens

y #mg B_2 per gram of turnip greens	x_1 radiation in gram calories per min.	x_2 avg. soil moisture tension	x_3 temperature (Fahrenheit)
110.4	1.76	0.07	7.8
102.8	1.55	0.07	8.9
101.0	2.73	0.07	8.9
108.4	2.73	0.07	7.2
100.7	2.56	0.07	8.4
100.3	2.80	0.07	8.7
93.7	1.84	0.07	8.7
96.6	1.98	0.02	7.6
96.2	0.80	0.02	6.7
99.0	0.80	0.02	6.2
88.4	1.05	0.02	7.0
75.3	1.80	0.02	7.3
77.1	2.30	0.02	8.2
65.7	1.91	0.47	8.3
56.8	1.91	0.47	8.2
62.1	1.91	0.47	6.9
53.2	2.13	0.47	7.6
59.4	2.13	0.47	6.9
58.7	1.51	0.47	7.5
58.0	2.05	0.47	7.6
102.8	2.80	0.07	7.4
98.9	2.16	0.07	8.8
99.4	0.59	0.02	6.5
92.0	1.80	0.02	6.5
82.4	1.77	0.02	7.6
74.0	2.03	0.47	7.6
61.0	0.76	0.47	7.4

TABLE 1.2. Regression Parameter Estimates for Model (1.7)

Term	Estimate	Std. Error	t Ratio	Prob > |t|
Intercept, b_0	82.4404	20.4066	4.04	0.0005
Radiation, b_1	2.6213	4.7441	0.55	0.5961
Moisture, b_2	-75.6944	39.6393	-1.91	0.0693
Temp, b_3	1.5206	3.1331	0.49	0.6322
Interaction, b_4	-1.6421	21.5369	-0.08	0.9339

TABLE 1.3 Summary of Fit for Estimated Model (1.7)

Number of Observations	27
Mean of Response Variable	84.233
R^2	0.754
R_a^2	0.709
Root MSE	10.131

to the same set of data, we obtain an R^2 of .9038, a root mean square error of 6.1958, and a set of predictor variables which are each statistically significant at a 0.05 level of significance. These results lead to the following questions:

1) How do we determine the "best" set of variables to be included in the model?
2) How can we know the correct degree and form of the functional model without making preliminary assumptions?
3) How can we ensure that the model obtained is, in fact, the "best approximating model" in terms of goodness of fit and complexity?

Given any list of potential independent variables, the investigator is usually responsible for screening out some of the variables based on the knowledge that a variable is not relevant to describing or predicting the behavior of the system or that two or more variables provide redundant information with regards to the response variable. Some potential variables are screened out because they are highly intercorrelated with other independent variables. Consider models (1.13) and (1.14) for the turnip green data in Table 1.1. Model (1.14) does not include x_1 in its functional

form. The correlation matrix provided in Table 1.4 shows a somewhat strong correlation between the variables x_1 and x_3. Due to this correlation between x_1 and x_3, the presence of both variables in the model may not add any information to the predictive capability of the model, but may increase the variation in estimation. Even with the aid of modern statistical tools, the variable selection process requires much subjective judgment. Which variable, x_1 or x_3, should be included in the model to "best" fit the turnip green data? How did Wakely (1949) determine the three independent variables used in our study out of all possible factors which potentially impact the amount of vitamin B_2 in turnip greens?

Criteria that have traditionally been used in variable selection procedures are R^2, as defined in (1.11), MSE, defined in (1.10), and Mallow's (1966) C_p. The C_p criterion is concerned concerned with the total mean squared error of the n fitted values for each of the models estimated from a subset of the set of predictive variables. For a full definition of the C_p criterion, see Neter, Wasserman, and Kutner

TABLE 1.4. Correlation Matrix for Predictor Variables of Turnip Green Data

Variable	Radiation, x_1	Moisture, x_2	Temperature, x_3
Radiation, x_1	1.000	0.012	0.537
Moisture, x_2	0.012	1.000	-0.014
Temperature, x_3	0.537	-0.014	1.000

(1983). As previously stated, maximizing R^2 does not take into account the number of parameters in the resulting model. Minimizing MSE, which adjusts R^2 by dividing by the degrees of freedom, is equivalent to maximizing the adjusted coefficient of multiple determination, defined in equation (1.6). Mallow's C_p assumes that MSE for the full model, with all independent variables included in the model, is a good estimate of the error variance. Also, Mallow's C_p cannot be used as a general criterion except for the linear regression-type models. Therefore, this thesis studies the use

of information-theoretic criteria as a more general means of variable selection that will reduce the need for subjective judgment.

For the data of Table 1.1, two different functional forms are specified in models (1.13) and (1.14). Each of these two forms were specified prior to the actual modeling of the data. For model (1.13) of the turnip data, the estimated model was found to be:

$$y = 82.4404 + 2.6213x_1 - 75.6944x_2 + 1.5206x_3 - 1.6421x_1x_2. \quad (1.15)$$

Estimation of (1.14) by the method of least squares yields the following estimated model:

$$y = 121.1013 + 496.188x_2 - 5.7981x_3 - 1129.529x_2^2 \quad (1.16)$$

The estimated model (1.16) yields a smaller mean square error than (1.15) and a higher R^2. Also, model (1.14) has fewer parameters to be estimated. Obviously, if models (1.13) and (1.14) are competing models, an investigator would clearly choose model (1.14) as the "best approximating model" to the data. This decision is made by

comparison of only two competing models. Obviously, there are many more competing models besides the two discussed in this example. For large, complex systems, it is onerous, if not impossible, to list, fit, evaluate, and compare all possible competing models. In our research, we propose a method that makes it possible to fitting a model to observed data without specifying the functional form of the model prior to the actual model development.

Evaluation of an estimated model typically includes testing the significance of the parameter estimates using the F test statistic and evaluating the strength of the relationship between the response variable and the predictor variables using R^2. These evaluation methods, as well as other model selection and evaluation criteria, are usually based on inferences about population means from normal populations. Thus, the criteria traditionally used to evaluate a model assume that the normal error model, defined in Section 1.3.2, is applicable.

1.5 Objective and Overview of Thesis

The overall objective or purpose of this thesis is to study some of the major problems which still exist in statistical modeling in general and to propose a new approach which addresses some of the problems by using for the first time Ivakhnenko's (1966) Group Method of Data Handling (GMDH) with information-theoretic model evaluation criteria. Therefore, this thesis develops the theory for the use of the information-based statistical criteria within the GMDH algorithm and it validates the approach through applications and Monte Carlo simulations.

In this chapter, we discussed basic notation, definitions, and preliminary ideas that we will need throughout the balance of the thesis. Chapter 1 also presented the general form of the model, as well as the data structure for our predictive model of interest, that we will be interested in throughout the study.

Chapter 2 presents and discusses Ivakhnenko's (1966) algorithm which is generally referred to as the *Group Method of Data Handling* (GMDH). In this chapter, the original, deterministic algorithm is defined, model evaluation methods which have been

used within the algorithm are discussed, and a numerical example of the algorithm is given. We provide a brief discussion of the development and history of the GMDH within the modeling framework. Finally, Chapter 2 establishes basic definitions and notations, common to Ivakhnenko's polynomial theory, which will be used throughout the thesis.

In Chapter 3, we briefly review the general theory of information-based statistical criteria for model identification and evaluation. We present the entropy-maximization principle, which serves as a premise for the information-based criteria. Also, in Chapter 3, the specific information-based criteria which we will use within the GMDH algorithm are defined and discussed within the framework of statistical modeling. Therefore, from a theoretical point of view, we introduce and discuss *Akaike's* (1973) *Information Criterion* (AIC) which provided the foundation for a modern approach to statistical model identification and evaluation based on information-theoretic statistics. Finally, we define Bozdogan's (1987, 1990) information-based criteria, ICOMP and ICOMP(IFIM), which incorporate a quantification of overall model complexity in the model evaluation criteria.

One of the major problems associated with traditional statistical modeling techniques is the assumption of normality of the error terms. The approach presented in this thesis allows for the modeler to assume different probability distributions for the random error terms. Therefore, in Chapter 4, we derive the information-based criteria for use within the GMDH algorithm under various probabilistic assumptions. Also, we establish the consistency of ICOMP and ICOMP(IFIM) as derived for the GMDH algorithm.

Chapter 4 also presents the new approach -- the GMDH algorithm using information-based criteria. We present the general form of the algorithm. This chapter also discusses how our approach deals intrinsically with some of the basic issues in statistical modeling, such as subset selection of variables and the existence of multicollinearity among the independent variables. We provide a numerical example to illustrate the stages of the algorithm and how the algorithm differs from traditional statistical modeling methods.

In Chapter 5, we present the use of a genetic algorithm within the GMDH framework. One of the proposed uses of the genetic algorithm is at each level of the algorithm to identify which models to pass on to the next generation of the algorithm. We also use a

genetic algorithm on the "best" model chosen by the GMDH algorithm to identify the variables to maintain in the "best" approximating model. After providing numerical examples using the genetic algorithm, we discuss both the strengths and limitations of the approach.

In Chapter 6, we validate our proposed approaches via Monte Carlo simulations. In our validation efforts, we compare the effectiveness of ICOMP and ICOMP(IFIM) as model evaluation statistics within the GMDH under the various probabilistic assumptions. Also, we compare the use of information-based statistics with other criteria that have been propose for use within the GMDH algorithm. Finally, we use simulations to study the effectiveness of our statistical modeling methodology under various situations, such as for times series data and when our response variable is an indicator variable.

In Chapter 7, we discuss future research topics within the framework of the GMDH algorithm and, in the Appendix, we provide the major computer programming modules for the GMDH algorithm.

CHAPTER 2

OVERVIEW OF THE GROUP METHOD OF DATA HANDLING (THE GMDH ALGORITHM)

Statistical modeling includes the identification of an appropriate model for a given realization, estimation of the values of the parameters of the identified model, validation of the model, and the final application of the model. As illustrated in Figure 1.2, these components are performed separately and sequentially. The proposed modeling technique integrates these components so that the identification of the model, the estimation of parameter values, and the validation of the model are treated as one entity and interactive throughout the model building process. This proposed method extends the GMDH algorithm. In this chapter, the general GMDH algorithm is presented and the research issues related to this algorithm are discussed.

2.1 Introduction

The group method of data handling (GMDH) is a combinatorial heuristic, developed by A. G. Ivakhnenko (1966), a Ukrainian cyberneticist, which constructs a mathematical model of a system in an evolutionary fashion. The algorithm is designed to model the deterministic component of complex systems. It constructs high-order regression type models beginning with a few basic quadratic equations and constructing from these equations new generations of more complex equations. The survival-of-the-fittest principle is used to determine which equations at each generation survive and get combined to form a new generation. As this algorithm continues for several generations, a set of mathematical models is developed that behave more and more like the actual system of interest (Farlow, 1981). As opposed to assuming a functional form of the model prior to parameter estimation, the GMDH algorithm considers only the data collected. The functional relationship between the response and predictor variables is learned directly from a self-organization of the data. When the number of predictors, m, is large, traditional modeling methods quickly become ill-behaved and

computationally unmanageable due to the large number of parameters that are simultaneously estimated. Ivakhnenko's algorithm avoids these problems by providing a systematic method of finding the "best" second degree polynomials, using these to create the "best" fourth degree polynomials, etc.

2.2 Description of Original Group Method of Data Handling (GMDH)

The GMDH algorithm is a heuristic developed to identify a mathematical model of a complex system by looking only at the data collected from the system and nothing else. The algorithm constructs a high-order polynomial of the Kolmogorov-Gabor form,

$$y = a + \sum_{i=1}^{m} b_i x_i + \sum_{\substack{i=1 \\ i \neq j}}^{m} \sum_{j=1}^{m} c_{ij} x_i x_j + \sum_{\substack{i=1 \\ i \neq j}}^{m} \sum_{\substack{j=1 \\ j \neq k}}^{m} \sum_{k=1}^{m} d_{ijk} x_i x_j x_k + \ldots, \quad (2.1)$$

by a composition of lower-order polynomials via a generative, self-organizing algorithm. These lower-order polynomials each take on the following general form:

$$z = A + Bx + Cy + Dx^2 + Ey^2 + Fxy. \quad (2.2)$$

Therefore, the GMDH algorithm combines second-order regression-type polynomials at each generation to arrive at the next generation of equations. The algorithm uses a data structure like that of multiple regression analysis. In the original algorithm as presented by Ivakhnenko and as usually applied, the data set is subdivided into two subsets, the *training set* and the *checking set*, for the purpose of cross-validation. This data structure is shown in Figure 2.1. We note that a comparison of Figure 2.1 with the data structure for our predictive model, shown in Figure 1.1, shows that one of the objectives of the proposed research is to eliminate the need for the subdivision of the data into training and checking sets.

Given the input data, each generation of the algorithm results in a set of quadratic polynomials with the inputs into the generation being polynomials from the previous layer. To find the "best" polynomials to fit the data, the GMDH algorithm passes through three major steps at each generation. These three primary steps are:

(1) Construction of new variables

(2) Screening of variables

(3) Optimality test

The following section explains each of these steps from Ivakhnenko's initial algorithm based on Farlow's (1981) summary and description.

2.2.1 The Steps of the GMDH Algorithm

Given a set of data of the form shown in Figure 2.1, the GMDH algorithm begins with the original m independent variables, x_1, x_2, ..., x_m, and the response variable, y. At each generation, the algorithm

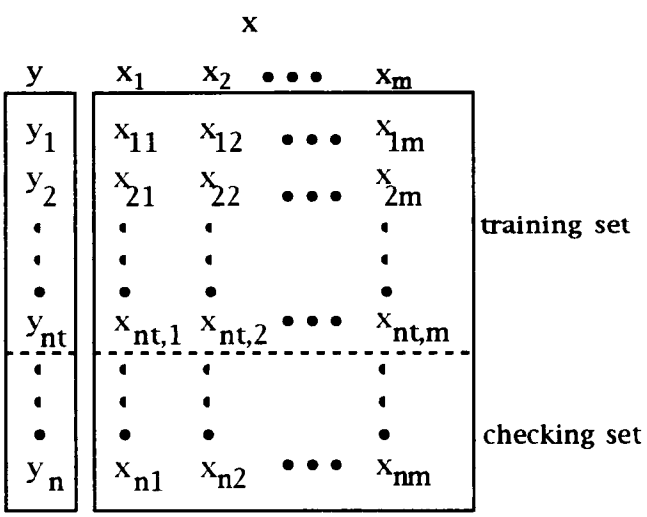

FIGURE 2.1 Data Structure of Original GMDH Algorithm

progresses through three primary steps. If the algorithm is at generation k, then the steps are as follows:

Step 1: (Construction of New Variables)

For each pair of indendent variables, construct the second-degree polynomials of the form shown in equation (2.2). We assume that there are m_k independent variables that entered into generation k. The number of polynomials of the form (2.2) constructed at generation k is equivalent to the number of pairs of the m_k independent variables, which is:

$$\frac{m_k(m_k - 1)}{2} \tag{2.3}$$

Thus, at generation one, the algorithm generates m(m-1)/2 second-degree polynomials. This construction of new higher-order variables (i.e., polynomials of degree two) is illustrated in Figure 2.2.

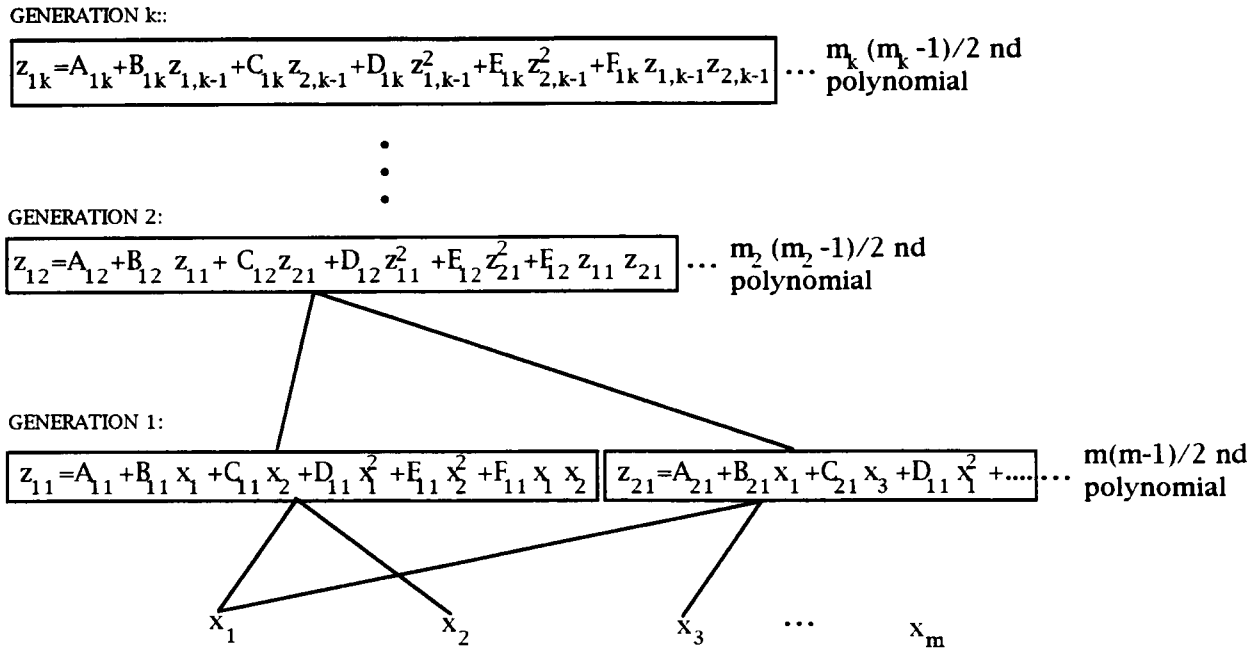

FIGURE 2.2. Computation of New Variables at Each Generation

Step 2: (Screening of Variables)

At the completion of step one, the algorithm has constructed a group of $m_k(m_k-1)/2$ new variables, $z_{1k}, z_{2k}, ..., z_{m_k k}$. These new variables replace the old predictor variables from the previous generation in the checking set. Thus, if k=1, the variables $z_{11}, z_{21}, ..., z_{m(m-1)/2,1}$ replace the original independent variables, $x_1, x_2, ..., x_m$, in the training set. Each of these new variables are evaluated for passage

into the next generation by determining which variables best estimate the dependent variable, y, in the checking set. In the algorithm as proposed by Ivakhnenko in 1966, the criterion used to evaluate the new variables at each generation is the *root mean square*, which is referred to by Ivakhnenko (1966) as the *regularity criterion*. The root mean square, denoted as r_j, is defined as:

$$r_j^2 = \frac{\sum_{i=nt+1}^{n} \left(y_i - z_{ij}\right)^2}{\sum_{i=nt+1}^{n} y_i^2}, \text{ for } j = 1, 2, \ldots, \binom{m_k}{2} \qquad (2.4)$$

From (2.4), we see that the regularity criterion is computed based on the values of the observations in the checking set. Also, with the use of the regularity criterion, an arbitrary "cut-off" value, R, is chosen such that for any $r_j > R$, the new variable is screened out and it is not passed on to the next generation of the algorithm.

Step 3: (Test for Optimality)

At the end of the second step, a value of r_j is obtained for each of the $m_k(m_k-1)/2$ variables in the kth generation. In step 3, the minimum value of these r_j's is determined and denoted as $RMIN_k$. If $RMIN_k$ is greater than RMIN, then the algorithm stops and the polynomial (i.e., the new variable) with the minimum value of the regularity criterion is chosen to be the "best" approximating model. Otherwise, $RMIN=RMIN_k$, and the algorithm moves on to the next generation and repeats these steps. According to Farlow (1984), it has been shown that the RMIN curve has the shape as shown in Figure 2.3. Thus, the algorithm does converge to a minimum value of the stopping criterion.

At the end of the GMDH algorithm, the final predicted values of the original response variable, y, are stored in the matrix Z. To determine the estimated coefficients in the higher-order polynomial, of the form shown in (2.1), the algorithm is backtraced and the second-order polynomials from each iteration are evaluated unitl a polynomial in the original variables (x_1, x_2, ..., x_m) is obtained.

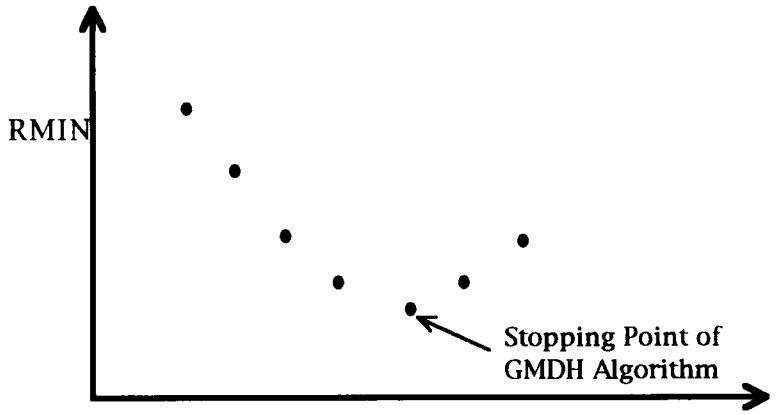

FIGURE 2.3. Behavior Curve of RMIN

2.3 A Simple Numerical Example of the Original GMDH Algorithm

We illustrate the functioning of the original GMDH algorithm using the Wakely (1949) data used in Table 1.1 of Chapter 1. In order to obtain a data structure of the form shown in Figure 2.1, the original data is subdivided into a training set and a checking set. This subdivision of the data is shown in Table 2.1. The algorithm begins with the original three (3) independent variables,

x_1 = radiation in gram calories per minute during preceding 1/2 day of sunlight,

x_2 = average soil moisture tension,

x_3 = air temperature in degrees Fahrenheit,

and the response variable,

y = milligrams of vitamin B_2 per gram of turnip greens.

Using this data, we progress through each of the three steps of the algorithm. The Matlab® computer modules which we developed for Ivankhenko's original algorithm are provided in Appendix B.

In step one of the algorithm in generation one, the second degree polynomials of the form in (2.2) are constructed using the training set of 14 observations. In generation one, there are three polynomials constructed. For the subdivision of data given in Table 2.1, the three polynomials in two variables constructed in level one are:

$$\begin{aligned} y_1 &= 104.82 - 29.99x_1 + 521.13x_2 + 5.97x_1^2 - 1240.93x_2^2 + 17.02x_1x_2 \\ y_2 &= 1614.6 + 63.26x_1 - 407.42x_3 + 3.29x_1^2 + 26.69x_3^2 - 7.99x_1x_3 \\ y_3 &= 369.85 + 476.74x_2 - 67.41x_3 - 1302.06x_2^2 + 3.7x_3^2 + 13.94x_2x_3 \end{aligned} \qquad (2.5)$$

TABLE 2.1. Subdivision of Wakely (1949) Data into Training and Checking Sets

Observation	y	x_1	x_2	x_3
		TRAINING SET		
1	110.4	1.76	.07	7.8
3	101.0	2.73	.07	8.9
5	100.7	2.56	.07	8.4
7	93.7	1.84	.07	8.7
9	96.2	0.80	.02	6.7
11	88.4	1.05	.02	7.0
13	77.1	2.30	.02	8.2
15	56.8	1.91	.47	8.2
17	53.2	2.13	.47	7.6
19	58.7	1.51	.47	7.5
21	102.8	2.80	.07	7.4
23	99.4	0.59	.02	6.5
25	82.4	1.77	.02	7.6
27	61.0	0.76	.47	7.4
		CHECKING SET		
2	102.8	1.55	.07	8.9
4	108.4	2.73	.07	7.2
6	100.3	2.80	.07	8.7
8	96.6	1.98	.02	7.6
10	99.0	0.80	.02	6.2
12	75.3	1.80	.02	7.3
14	65.7	1.91	.47	8.3
16	62.1	1.91	.47	6.9
18	59.4	2.13	.47	6.9
20	58.0	2.05	.47	7.6
22	98.9	2.16	.07	8.8
24	92.0	1.80	.02	6.5
26	74.0	2.03	.47	7.6

These three polynomials, y_1, y_2, and y_3, replace the original set of predictor variables for the algorithm.

In step two of generation one, the three new variables in (2.5) are screened by comparing the computed regularity criterion (2.4) to the user's chosen cut-off value, R. In this initial example, we have chosen a cut-off value as R=0.40. The value of the regularity criteria for each new variable (i.e., model constructed) is obtained using the checking set of 13 observations. The values of the regularity criterion for each of the three models in level one are:

New Variable	Regularity Criterion (r_j)
y_1	0.1040
y_2	0.2651
y_3	0.0914

Since, for each of the new variables (i.e., models) constructed in level one, it is the case that $r_j < R$, then no model is screened out and all pass to the second level of the algorithm. In step three, the minimum value of the regularity criterion for the three models (2.5)

is determined. $RMIN_1 = 0.0914$ for y_3. Since this is the first generation of the algorithm, the overall minimum value of the regularity criterion, RMIN, is equal to the minimum value for the first generation. So, $RMIN = RMIN_1 = 0.0914$.

The models generated in step one of generation two of the algorithm, using y_1, y_2, and y_3 as the predictor variables and 14 observations in the training set, are:

$$\begin{aligned}
\bar{y}_1 &= 33.642 + 2.21y_1 - 1.96y_2 + 0.002y_1^2 + 0.021y_2^2 - 0.02y_1y_2 \\
\bar{y}_2 &= 40.227 + 4.224y_1 - 4.282y_3 - 0.37y_1^2 - 0.32y_2^2 + 0.696y_1y_3 \qquad (2.6) \\
\bar{y}_3 &= 77.43 - 2.84y_2 + 2.12y_3 + 0.023y_2^2 - 0.013y_2y_3
\end{aligned}$$

The computed values of the regularity criterion for these new variables in (2.6) are:

New Variable	Regularity Criterion (r_i)
$\bar{y}_1$	0.1685
$\bar{y}_2$	0.3840
$\bar{y}_3$	0.1761

Again, all of the regularity criterion values are less than the cut-off value, R=0.40. Thus, we continue to step three of the algorithm, the "test for optimality". $RMIN_2 = 0.1685$ and RMIN = 0.0914. Hence, the algorithm stops and the polynomial with the minimum value of the regularity criterion, y_3 of (2.5), is chosen to be the "best" approximating model for the data. So, the chosen model using the original GMDH algorithm under the current subdivision of the data is:

$$y_3 = 369.85 + 476.74x_2 - 67.41x_3 - 1302.06x_2^2 + 3.7x_3^2 + 13.94x_2x_3 \quad (2.7)$$

For this same set of data, a different subdivision of the data is obtained through the use of a random number table. This new subdivision of the Wakely data into the training and checking sets is provided in Table 2.2. Given the new observations in the training set, the three polynomials generated in level one are:

TABLE 2.2. A Second Subdivision of Wakely (1949) Data into Training and Checking Sets

Observation	y	x_1	x_2	x_3
		TRAINING SET		
3	101.0	2.73	.07	8.9
6	100.3	2.80	.07	8.7
7	93.7	1.84	.07	8.7
9	96.2	0.80	.02	6.7
11	88.4	1.05	.02	7.0
13	77.1	2.30	.02	8.2
16	62.1	1.91	.47	6.9
18	59.4	2.13	.47	6.9
19	58.7	1.51	.47	7.5
21	102.8	2.80	.07	7.4
22	98.9	2.16	.07	8.8
24	92.0	1.80	.02	6.5
25	82.4	1.77	.02	7.6
26	74.0	2.03	.47	7.6
		CHECKING SET		
1	110.4	1.76	.07	7.8
2	102.8	1.55	.07	8.9
4	108.4	2.73	.07	7.2
5	100.7	2.56	.07	8.4
8	96.6	1.98	.02	7.6
10	99.0	0.80	.02	6.2
12	75.3	1.80	.02	7.3
14	65.7	1.91	.47	8.3
15	56.8	1.91	.47	8.2
17	53.2	2.13	.47	7.6
20	58.0	2.05	.47	7.6
23	99.4	0.59	.02	6.5
27	61.0	0.76	.47	7.4

$$
\begin{aligned}
y_1 &= 108.25 - 29.22x_1 + 284.87x_2 + 6.56x_1^2 - 812.22x_2^2 + 36.23x_1x_2 \\
y_2 &= 814.27 + 24.6x_1 - 208.78x_3 + 34.49x_1^2 + 16.73x_3^2 - 20.63x_1x_3 \\
y_3 &= 338.20 + 117.27x_2 - 60.28x_3 - 1043.59x_2^2 + 3.34x_3^2 + 47.64x_2x_3
\end{aligned}
\tag{2.8}
$$

The values of the regularity criterion for each of these models in (2.8) are:

New Variable	Regularity Criterion (r_i)
y_1	0.09185
y_2	0.2569
y_3	0.1169

At generation two of the algorithm, the $RMIN_2 = .0948$. Since $RMIN_1 < RMIN_2$, the algorithm stops and the chosen "best" approximating model is y_1 in (2.8).

2.4 Discussion of the Original GMDH Algorithm

The GMDH algorithm has many strengths which warrant its use and continued development. Even though the algorithm is a heuristic method (i.e., it has yet to be proven that the algorithm yields global optimum), there has been much evidence to show that the algorithm performs well while relaxing the assumption of a known functional form prior to the modeling process. Ikeda, Ochiai, and Sawaragi (1976) effectively used a form of the GMDH for river flow prediction. Rice (1979) showed the use of the algorithm for climate/crop forecasting. Kenichi Ohashi of The World Bank (1984) developed GMDH forecasting models for the forecasting of two key U.S. interest rates. Additionally, there have many more who have chosen to use the GMDH algorithm (or a variant of the GMDH) for the modelling of economic and scientific systems.

The reasons for the growing attention to the GMDH are primarily due to its ability to computationally develop a higher-order model without requiring a huge number of observations. If we believe that a system potentially has 8 predictor variables and has a complexity to at least the fourth order, then the analysis of the

regression model requires 495 parameters to be estimated. Therefore, to obtain the coefficient estimates, a minimum of 500 observations is necessary. Thus, with smaller data sets, the modeller can model higher-order systems without having specific knowledge of the expected functional form of the system.

There are, however, limitations to the GMDH as originally proposed by Ivakhnenko. These limitations are primarily in three areas -- the subdivision of the data, the use of the regularity criterion, and the tendency of the algorithm to overspecify the model. As seen in the previous example of Section 2.3, two different subdivisions of the data can yield very different results. For the first subdivision of the data, shown in Table 2.1, the algorithm chose a second-order model in x_2 and x_3. However, the second subdivision of the data (determined from generated random numbers) in Table 2.2 yielded a choice of model to be second-order in x_1 and x_2. Obviously, since each model was developed from only a portion of the data set, the information used to construct the model is different for each subdivision of data. The loss of information to use in modelling is a major weakness of the original algorithm, especially since the

algorithm is developed to allow for the modelling of large systems with small amounts of data. To be able to generate self-organized models from the data, we desire to use the most information as possible. Thus, the work described in the following chapters eliminates this need to subdivide the data.

The regularity criterion of (2.4) is the proposed criteria for making decisions regarding the models to pass on to the next generations as well as regarding when to stop the algorithm. This criterion is dependent on the subdivision of the data (as it is constructed on the checking set). It also has a major disadvantage in that it is a "poor selector in the presence of noisy data and in the problem of medium- and long-term prediction" (Farlow, 1984). Several other criteria, such as the unbiased criterion and the combined criterion, have been proposed to make the criteria less sensitive to noise. For references on these proposed variations to the regularity criterion, see Ivakhnenko and Ivakhnenko (1974) and Ikeda et al. (1976). Again, in this thesis, we make use of the properties of the information-theoretic class of statistical model

evaluation criteria to provide greater robustness to noisy data partially due to the elimination of the need for subdivision of data.

Finally, the GMDH algorithm by nature tends to overspecify the model for several reasons. First, the use of the usual form of the Ivakhnenko polynomial of (2.2) implies that the simplest model of interest is a second-order polynomial. After the second iteration, the degree is 4, then 8, and so on. Since the degree of the polynomial doubles at each generation, the algorithm will never inherently choose a third-order polynomial as the "best" approximating model. In order to overcome this weakness, we develop a genetic algorithm for use within the GMDH to determine the best models to pass on to each generation as well as the best terms to include within each model.

CHAPTER 3

GENERAL THEORY OF INFORMATION-BASED STATISTICAL CRITERIA FOR MODEL IDENTIFICATION AND EVALUATION

As we discussed in the previous chapter, most of the criteria used within the GMDH algorithm are poor selectors in the presence of noisy data. The regularity criterion (2.4) and variations of this criterion do not take model complexity into account. In this chapter, we present the most important theoretical aspect of this thesis. We, for the first time, propose the use of information-based statistical criteria within the GMDH algorithm. Thus, this chapter plays an important role between the previous introductory chapters and the following chapters in which we develop the algorithm using information-based criteria.

3.1 Introduction

In the following sections, we will develop the underlying theory of information-based statistical criteria. Despite the recent development of statistical modeling in many different disciplines, there still exists difficulty in the ability to construct an adequate model based on the available sample information. As we develop the theoretical background of the proposed criteria, we must remember that, even though the proposed techniques depend heavily on their theoretical bases, many theoretically sound statistical techniques and criteria have practical limitations and weaknesses.

3.1.1 Review of Conventional Procedures

As described by Figure 1.1, there is a clear breakdown of conventional modeling techniques into two primary, but separate, stages: estimation of parameters and evaluation of model through a type of hypothesis-testing problem. The theory concerned with the *estimation* of distribution functions and the parameters of models built around random variables is well established. This well known

theory of estimation was developed by R. A. Fisher (1912, 1921, 1925, 1934, and 1935).

The second part of most statistical modeling efforts is the *testing* of the validity of the hypotheses about the specified model or distribution function and the parameters of that function or model. The question to be answered is: "Is the observed data consistent with the stated hypotheses?" In this component of statistical modeling, most criteria are developed for use within the framework of the Neyman-Pearson theory of statistical hypothesis testing. These criteria are therefore based on two possibilities: the proposed hypothesis concerning the appropriateness of the fitted model is either accepted or rejected. If this proposed hypothesis is accepted and it is wrong, then an incorrect decision regarding the appropriateness of the estimated model is made. This type of incorrect decision is referred to as a *Type II error.* The probability of a Type II error is denoted by β. On the other hand, if the proposed hypothesis is rejected and it is correct, then another incorrect decision is made. This type of incorrect decision is referred to as a *Type I error.* The probability of a Type I error occurring is

typically denoted by α. Since both of these error types cannot be simultaneously controlled given a set of observations, it is usual in traditional evaluation methods to set α at a pre-assigned, arbitrary, small value. This is stated by Lehman (1959). "*The choice of a level of significance will usually be somewhat arbitrary since in most situations there is no precise limit to probability of an error of the first kind that can be tolerated. It has become customary to choose from one of a number of standard values such as .005, .01, or .05. There is some convenience in standardization since it permits a reduction in certain tables needed for carrying out certain tests. Otherwise, there appears to be no particular reason for selecting these values.*"

Information based statistical criteria provide a method by which the estimation and testing components of statistical modeling are performed simultaneously. They also eliminate the need for the arbitrary identification of a value for α as in traditional Neyman-Pearson forms of testing. Therefore, we look at the theory of information-based statistical criteria and propose their use in statistical modeling methods as a way to eliminate some of the

weaknesses of both the original GMDH algorithm and of the conventional theory of testing and estimation.

3.1.2 Definitions and Background of Information-Based Statistical Criteria

The introduction of *Akaike's* (1971) *information criterion* (AIC) provided the foundation for a modern approach to statistical model identification and evaluation based on information theory. Information theory is a branch of the mathematical theory of probability and statistics which is based on Boltzmann's (1877) concept of entropy. Akaike's development of the statistical theory of AIC was stimulated largely by the mathematical and statistical works of Fisher (1956), Shannon (1956), and Wiener (1956). Fisher first introduced the concept of information quantity in a statistical sense in 1925 in his work on the theory of estimation.

Following Akaike's initial introduction of his criterion, he continued to develop the theory behind his entropic, information-based criterion in a series of papers (1973, 1974, 1977, and 1981). *Information-based model selection criteria* are statistical criteria

used to choose the "best" model from a set of competing models using information quantity as a measure of goodness-of-fit. *Entropy* is empirically defined to be a measure of uncertainty in a set of collected data. If the entropy in the range of data values is large, the information provided in the set of data will be small. The AIC criterion provides a sample estimate of the expected entropy and is a direct extension of Boltzmann's generalized entropy and Kullback and Leibler's (1951) information-theoretic interpretation of the method of maximum likelihood. Most information-based statistical criteria extend from Akaike's work and measure the closeness of the approximation between the fitted model and the true model based on the concept of entropy.

3.1.3 The Concept of Entropy

A model can be expressed as an estimation of the true probability distribution function based on the set of data. Given this expression, an effective model selection criterion minimizes the distance between the model and the true distribution. In other

words, the criterion selects the model with the most information about the true distribution. There exist two measures based on this concept: Boltzmann's generalized entropy and the Kullback-Leibler information quantity. As discussed in the previous section, the major development of AIC lies in the extension of an information-theoretic interpretation of the method of maximum likelihood and the concept of entropy. Entropy is a natural measure of discrimination between the true distribution and the estimated or fitted model.

Let $f(x|\theta)$ represent the true distribution, where x is a continuous random vector of interest and θ is the vector of parameters. Also, let $g(x|\hat{\theta})$ represent the model estimating $f(x|\theta)$. Boltzmann's generalized entropy is defined to be:

$$\begin{aligned} B\,[f(x\mid\theta); g(x\mid\hat{\theta})] &= E[\log g(x\mid\hat{\theta}) - \log f(x\mid\theta)] \\ &= \int f(x\mid\theta)\log g(x\mid\hat{\theta})dx - \int f(x\mid\theta)\log f(x\mid\theta)dx \\ &= H[f(x\mid\theta);\ g(x\mid\hat{\theta})] - H[f(x\mid\theta);\ f(x\mid\theta)], \end{aligned} \tag{3.1}$$

where $H[f(x\mid\theta);\ g(x\mid\hat{\theta})]$ is the cross-entropy that determines the goodness-of-fit of $g(x\mid\hat{\theta})$ to $f(x|\theta)$. Now, K-L information is defined to be:

$$I[f(x \mid \theta);\ g(x \mid \hat{\theta})] = H[f(x \mid \theta);\ f(x \mid \theta)] - H[f(x \mid \theta);\ g(x \mid \hat{\theta})]$$
$$= -B[f(x \mid \theta);\ g(x \mid \hat{\theta})]. \qquad (3.2)$$

Some of the analytic properties of $I[f(x \mid \theta);\ g(x \mid \hat{\theta})]$ from Kullback (1959) are:

(1) If $f(x \mid \theta) \neq g(x \mid \hat{\theta})$, then $I[f(x \mid \theta);\ g(x \mid \hat{\theta})] > 0$.

(2) If $f(x \mid \theta) = g(x \mid \theta)$, then $I[f(x \mid \theta);\ g(x \mid \hat{\theta})] = 0$.

(3) If $x_1, x_2, ..., x_n$ are independent, identically distributed random variables, then the K-L information for the entire sample is $I_n[f(x \mid \theta);\ g(x \mid \hat{\theta})] = nI[f(x \mid \theta);\ g(x \mid \hat{\theta})]$.

Because of the duality in (3.2) between Boltzmann's entropy and the K-L information, we can minimize the K-L information, which is derivable from minimal assumptions, instead of maximizing Boltzmann's entropy given in (3.1). The K-L information depends on the true, unknown distribution and parameters. Thus, in order to be able to use K-L information to evaluate and compare various models, this quantity must be estimated from the observed data.

The K-L information can be estimated from the observed data through its relationship to the mean or expected log likelihood. Again, if $f(x|\theta)$ is the true model and $g(x|\hat{\theta})$ is the entertained model, then we can express the K-L information as:

$$I[f(x|\theta);\ g(x|\hat{\theta})] = \frac{1}{n}E[\log L(x|\theta) - \log L(x|\hat{\theta})],\ \text{or}$$

$$nI[f(x|\theta);\ g(x|\hat{\theta})] = E[\log L(x|\theta) - \log L(x|\hat{\theta})], \qquad (3.3)$$

where $L(\theta)=f(x_1, x_2, ..., x_n|\theta)$ is the likelihood function for the set of data as defined in (1.3). The average log likelihood, $1/n \log L(\theta)$, can be interpreted as an estimator of the distance between $f(x|\theta)$ and $g(x|\hat{\theta})$, the true and entertained models, respectively. Hence, we can use the expected log likelihood function, over all models, to estimate the K-L information.

3.1.4 The Entropy Maximization Principle

Given the definition of entropy in (3.1) and the relationship described in (3.3), we can state the *entropy maximization principle* according to Akaike (1977):

Formulate the object of statistical inference as the estimation of the true distribution $f(x|\theta)$ from the data $x=(x_1, x_2, ..., x_n)$ and try to search for an approximate model $g(x|\hat{\theta})$ which will maximize the expected entropy:

$$E_X[B(\theta;\hat{\theta})] = E_X\{E[\log\ g(x|\hat{\theta})] - E[\log\ f(x|\theta)]\} \qquad (3.4)$$

Since the second term in the right-hand side of (3.4) is a constant depending only on $f(x|\theta)$, we can consider only the first term,

$$E_X\{E[\log\ g(x|\hat{\theta})]\}, \qquad (3.5)$$

for the purpose of comparing fitted models. The quantity in (3.5) provides a reasonable criterion for defining a best fitting model by its maximization or by minimizing its negative. For more details, refer to Bozdogan (1987).

3.2 Information-Theoretic Model Evaluation Criteria: AIC, ICOMP, and ICOMP(IFIM)

In this section, we present AIC and other information-theoretic criteria in terms of their relationship to entropy and/or information in a set of data.

3.2.1 Akaike's Information Criterion (AIC)

Akaike's Information Criterion (AIC) is a sample estimate of the expected log likelihood or, equivalently, the expected K-L information in a set of data. Thus, AIC can be used to compare the models in a set of competing models, as stated in the following proposition:

Proposition: Let $\{m_k: k=1,2,...,K\}$ be a set of competing models indexed by k=1,2,...,K. Then, the criterion

$$AIC(k) = -2 \log L(\hat{\theta}_k) + 2k, \tag{3.6}$$

which is minimized to choose a model M_k over the set of competing models, is a sample estimator of $2E[I(\theta;\hat{\theta}_k)]$ or $-2E[\log\ g(x\,|\,\hat{\theta}_k)]$ of the true distribution with respect to a model with the parameters determined by the method of maximum likelihood. The proof of this proposition is provided by Bozdogan (1987).

From this proposition and Akaike's Entropy Maximization Principle (3.5), it is easily seen that AIC estimates E[nI]=-E[nB]. The form of AIC provided in (3.6) is readily interpreted as:

AIC = (-2)log(maximized likelihood) + 2(no. of free parameters in model).

The first term provides us with a measure of bias or model inaccuracy when the maximum likelihood estimators of the parameters in the model are used. The second term serves as a penalty for the increased unreliability for the bias in the first term when additional free parameters are included in the model. Since the decision rule is to minimize AIC over the set of competing models, the "best approximating" model is chosen to be the one with the highest information gain. The application of AIC emphasizes the

comparison of the goodness-of-fit of the competing models while also taking into account the principle of parsimony.

3.2.2 Bozdogan's ICOMP Criterion

Given that any model is an estimation or approximation to the true distribution of the system of interest, some error or loss is incurred in the modeling process. For a general linear or nonlinear structural model defined by

$$\text{Statistical Model} = \underbrace{\text{Signal}}_{\text{Deterministic Component}} + \underbrace{\text{Noise}}_{\text{Random Component}},$$

the overall loss due to modeling can be expressed as:

Loss = Badness of Fit + Lack of Parsimony + Profusion of Complexity.

As seen in (3.6), AIC estimates badness of fit by minus twice the expected log likelihood and it uses 2k, where k represents the number of free parameters in the model, to penalize a model for lack of parsimony. Several authors, including Rissanen (1976), have questioned whether 2k is a sufficient penalty term in order to

prevent over-specialization and unnecessary complexity by the chosen model.

Motivated by the work of Akaike, Bozdogan (1988, 1990) developed an informational criterion, ICOMP, which incorporates a measure of the complexity of the model into the penalty term. Empirically, *complexity* is defined to be a measure of the interdependence among the components of a random vector. Since large values of complexity indicate a high interaction among the variables and smaller values of complexity indicate less interaction among the variables, the concept of complexity plays an important role in model evaluation. Theoretically, to define an information-theoretic measure of complexity, we use the concept of entropy provided in Section 3.1.3 and the relationship in Shannon's (1948) entropy identity:

$$I(X) \equiv I(x_1, x_2, ..., x_q) = \sum_{j=1}^{q} H(x_j) - H(x_1, x_2, ..., x_q), \qquad (3.7)$$

where $H(x_j)$ is the marginal entropy of x_j and $H(x_1,x_2,...,x_q)$ is the joint entropy among the set of random variables. Based on this

relationship in (3.7), Van Emden (1971) provides a reasonable initial definition of informational complexity of a random vector with covariance matrix Σ:

$$C_0(\Sigma) = \frac{1}{2}\left[\sum_{j=1}^{q} \log(\sigma_j^2) - \log|\Sigma|\right], \tag{3.8}$$

where σ_j^2 is the variance of the x_j.

As discussed by Van Emden (1971), the original covariance complexity (3.8) depends upon both the marginal and joint distributions of the random variable and is not invariant under orthonormal transformations. Therefore, Bozdogan (1990) proposed the use of a maximal information theoretic measure of complexity to characterize the covariance matrix properties of the parameter estimates and the error terms. From Bozdogan (1990) and Van Emden(1971), the maximal complexity of a random vector with positive definite covariance matrix Σ is defined as:

$$C_1(\Sigma) = \frac{q}{2}\log\operatorname{tr}\left(\frac{\Sigma}{q}\right) - \frac{1}{2}\log(\det(\Sigma)), \tag{3.9}$$

where q is the dimension or rank of Σ. $C_1(\Sigma)$ measures both inequality among variances as well of the contribution of the covariances. It has been shown (Bozdogan, 1990) that $C_1(\Sigma)$ is invariant under orthonormal transformations.

ICOMP, as an information-theoretic model evaluation criterion for linear or nonlinear statistical models, uses an information-based characterization of the covariance matrix properties of the parameter estimates and the error terms starting from their finite sampling distributions. ICOMP is defined to be:

$$\text{ICOMP} = -2 \log L(\hat{\theta}) + 2C_1(\hat{\Sigma}(\hat{\theta})) + 2C_1(\hat{\Sigma}(\hat{\varepsilon})), \qquad (3.10)$$

where $2C_1(\hat{\Sigma}(\hat{\theta}))$ represents the maximal complexity (3.9) due to the estimated covariance structure of the parameter estimates and $2C_1(\hat{\Sigma}(\hat{\varepsilon}))$ represents the complexity due to the covariance matrix of the residuals.

In (3.10), we see that, in addition to the first term that provides an estimate of badness of fit, ICOMP also incorporates a measure of complexity due to the covariance matrix of the estimated

parameters and a measure of complexity due to the covariance of the error terms. It controls the loss due to insufficient and overparameterized models, and it incorporates the assumption of dependence of residuals in a single model evaluation criterion. If the error terms are uncorrelated, then the last term of (3.7), $C_1\left(\hat{\Sigma}(\hat{\varepsilon})\right)$, tends to zero. Among a portfolio of competing models, the model with the minimum value of ICOMP is chosen to be the "best" model. Thus, the "best" model is the one that achieves the most sufficient trade-off between the accuracy of the parameter estimates and the model complexity.

3.2.3 Bozdogan's ICOMP(IFIM) Criterion

A second approach to quantifying the concept of overall model complexity is to use the estimated inverse-Fisher information matrix (IFIM), also known as the Cramer-Rao lower bound matrix. In this case, Bozdogan (1990, 1994) defines the information-theoretic model evaluation criterion to be:

$$\text{ICOMP(IFIM)} = -2 \log L\left(\hat{\theta}\right) + 2C_1\left(\hat{F}^{-1}\right), \qquad (3.11)$$

where C_1 denotes the maximal information complexity as defined in (3.9) and $\hat{F}^{-1}$ represents the inverse Fisher information matrix. In (3.11), the inverse-Fisher information matrix evaluated at $\hat{\theta}$ is used to measure the complexity of the parameter estimates.

This approach to model evaluation takes into account the accuracy of the estimated parameters and implicitly adjusts for the number of free parameters in the model. If $F_{jj}^{-1}(\theta)$ represents the diagonal elements of the inverse Fisher information matrix (j=1,...,q), then we know from Chernoff (1956) that these elements represent the variance of the asymptotic distribution of $\sqrt{n}(\hat{\theta}_i - \theta_i)$ for i=1,...,q. Now, consider a subset of the q parameters of size k. Then, we have that

$$F_{jj}^{-1}(\theta_q) \geq F_{jj}^{-1}(\theta_k). \qquad (3.12)$$

Behboodian (1964) explains that the inequality in (3.12) means that the variance of the asymptotic distribution of $\sqrt{n}(\hat{\theta}_i - \theta_i)$ can only increase as the number of unknown parameters is increased. Thus, the use of $C_1(\hat{F}^{-1}(\hat{\theta}))$ in the information-theoretic model evaluation criterion defined in (3.10) takes into account the fact that as we increase the number of free parameters in a model, the accuracy of the parameter estimates decreases. Hence, ICOMP(IFIM) chooses simpler models that provide more accurate and efficient parameter estimates over more complex, overspecified models as preferred according to the principle of parsimony.

3.3 Conclusion

In order to use information-theoretic criteria for model identification and evaluation, we must first have a probabilistic structure of a model and a defined likelihood function of the model. For all of the discussed criteria, when several competing models exist, the procedure is to estimate the parameters within the models, compute the value of the information-theoretic criteria, and compare the models choosing the model with the minimum value of the

criterion to be the "best". Practically, these criteria provide a simple, yet effective means of model evaluation and comparison without the need to set arbitrary levels of significance and without needing to look up any tabulated values. The effectiveness of these criteria are directly related to their theoretical justifications. Thus, we will derive the information-theoretic criteria for the GMDH algorithm under the assumption of normality and discuss their theoretical justifications in the following chapter. We will also provide a numerical example and show some empirical results.

CHAPTER 4

INFORMATION-BASED CRITERIA WITHIN THE GMDH ALGORITHM UNDER THE ASSUMPTION OF NORMALITY

In order to implement and evaluate the proposed improvement of the traditional Group Method of Data Handling, an elegant approach to statistical data modeling, we must derive the information-based model evaluation criteria for the GMDH. In the initial derivation of these criteria, the following assumptions hold:

(1) The probability distributions of the response variable have the same variance, σ^2, regardless of the levels of the independent or predictor variables.

(2) The error terms and, hence, the response variables are normally distributed.

Thus, in this chapter, the proposed statistical modeling and model evaluation technique allows for the relaxation of the following two

assumptions associated with many of the traditional statistical modeling techniques:

(1) The functional form of the model is known prior to the modeling of the data.

(2) The error terms are uncorrelated.

4.1 Introduction

Recall the form of the lower-order polynomials given in (2.2) constructed in each stage (i.e., generation) of the GMDH algorithm:

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_i x_j, \text{ for } \forall(i, j).$$

Under the assumptions that y is normally distributed and that there exists, therefore, a random error component additive with the above deterministic component, the polynomials generated at each level of the GMDH algorithm can be modeled as regression-type equations of the following form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2, \qquad (4.1)$$

where the random vector, y, has mean or expectation given in matrix form:

$$E(y) = X\beta, \tag{4.2}$$

and ε is a vector of normal random variables with mean 0 and variance-covariance matrix $\sigma^2 I$. We can analytically express the density function for a particular sample observation as:

$$f(y_i \mid x_i\beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(y_i - x_i'\beta)^2}{2\sigma^2}\right]. \tag{4.3}$$

The likelihood function of the sample is:

$$\begin{aligned} L(\beta, \sigma^2 \mid x_i\beta, \sigma^2) &= \prod_{i=1}^{n} f\ (y_i \mid x_i\beta, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{y'y - 2\beta'X'y + \beta'X'X\beta}{2\sigma^2}\right]. \end{aligned} \tag{4.4}$$

The maximum likelihood estimators of β and σ^2 are given by:

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{4.5}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(y - \hat{y})'(y - \hat{y}) = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon}. \tag{4.6}$$

4.2 Derivation of Information-Based Statistical Criteria for the GMDH Algorithm

Given the model in (4.1) under the assumption of normality and the definitions of AIC in (3.6), ICOMP in (3.10) and ICOMP(IFIM) in (3.11), we can use the maximum likelihood (ML) principle to obtain the first terms in each of the criteria. The extensive use and development of the maximum likelihood principle stems from the work of Fisher (1921). The principle basically states that, in order to estimate the parameter θ, we use that value (referred to as $\hat{\theta}$) which makes the likelihood function as large as possible. Thus, we use this principle to obtain parameter estimates in the derivation of the information-based criteria for the GMDH algorithm under the assumption of normality.

4.2.1 Derivation of AIC($GMDH_{normal}$)

Given the model in (4.1) and the definition of AIC provided in (3.6), we can maximize the log likelihood function of the model to obtain AIC for the GMDH algorithm:

$$\begin{aligned} AIC(GMDH_{normal}) &= -2\log L(\hat{\beta},\hat{\sigma}^2)+2p \\ &= -2\left\{\left[-\frac{n}{2}\log(2\pi)\right]-\frac{n}{2}\log(\hat{\sigma}^2)-\frac{n}{2}\right\}+2p \\ &= n\log(2\pi)+n\log(\hat{\sigma})+n+2p, \end{aligned} \tag{4.7}$$

where p represents the number of parameters estimated in the model, or p-1 represents the number of predictors in the model. We note that, in the GMDH algorithm, the last term, 2p, which exists to penalize for lack of parsimony, is constant across the competing models at each generation of the algorithm. Thus, AIC($GMDH_{normal}$) does not evaluate competing models on the principle of parsimony or on complexity, but merely on goodness-of-fit. Since it is well known that increasing the number of parameters in a model improves most

measures of goodness-of-fit, AIC($GMDH_{normal}$) has a tendency to overfit in the GMDH environment. This tendency is seen in the first numerical example in Section 4.3. Thus, from this point forward, except for exemplary purposes, we will use ICOMP or ICOMP(IFIM) for model selection and evaluation in the GMDH algorithm.

4.2.2 Derivation of ICOMP($GMDH_{normal}$)

In order to take model complexity into account, we derive the ICOMP criterion (3.10) for the GMDH algorithm under the assumption $y \sim N(\mathbf{X}\beta, \sigma^2)$. Under this situation, ICOMP is defined by:

$$\begin{aligned} ICOMP(GMDH_{normal}) &\equiv ICOMP(\hat{\beta}, (\hat{\varepsilon} \mid \hat{\beta})) \\ &= -2\log L(\hat{\beta}, \hat{\sigma}^2) + 2[C_1(\hat{\Sigma}(\hat{\beta})) + C_1(\hat{\Sigma}(\hat{\varepsilon}))]. \quad (4.8) \end{aligned}$$

The first component in (4.8), $-2\log L(\hat{\beta}, \hat{\sigma}^2)$, is obtained by maximizing the log likelihood function. The second term of

ICOMP($GMDH_{normal}$) is obtained from the definition of maximal complexity (3.9):

$$C_1(\hat{\Sigma}(\hat{\beta})) = \frac{q}{2}\log\left[\frac{tr\hat{\sigma}^2(X'X)^{-1}}{q}\right] - \frac{1}{2}\log\left|\hat{\sigma}^2\,(X'X)^{-1}\right| \qquad (4.9)$$

Thus, we have that $C_1(\hat{\Sigma}(\hat{\beta})) = C_1(\hat{\sigma}^2(X'X)^{-1})$. Since maximal complexity is invariant under scalar multiplication, $\hat{\sigma}^2$ can be factored out. Thus, the second term of the ICOMP criterion under the assumption of normality becomes:

$$C_1(\hat{\Sigma}(\hat{\beta})) = C_1((X'X)^{-1}). \qquad (4.10)$$

The third component of ICOMP($GMDH_{normal}$) is the estimated maximal complexity of the variance-covariance matrix of the residuals. In order to obtain this component, we use the definition of maximal complexity (3.9) and the following relationship between the vector of residuals and the hat matrix:

$$\underset{nx1}{\varepsilon} = \underset{nxn}{(I - H)} \underset{nx1}{Y} \qquad (4.11)$$

where the hat matrix, H, is defined as:

$$H = X(X' X)^{-1}X' \qquad (4.12)$$

The matrix (I-H) has the properties of symmetry and idempotency. From these properties, the variance-covariance matrix of ε can be obtained from (4.11) as:

$$\Sigma(\varepsilon) = (I-H)\sigma^2(Y)(I-H)' \qquad (4.13)$$

For the normal error model, $\sigma^2(Y) = \sigma^2(\varepsilon) = \sigma^2 I$. Also, $(I-H) = (I-H)'$. Hence, (4.13) becomes:

$$\Sigma(\varepsilon) = \sigma^2(I - H). \qquad (4.14)$$

Substituting (4.14) into the definition of maximal complexity (3.9), we obtain the following general form for the third term:

$$C_1(\hat{\Sigma}(\hat{\varepsilon})) = \frac{n}{2}\log\left[\operatorname{tr}\left(\frac{\hat{\sigma}^2(I - H)}{n}\right)\right] - \frac{1}{2}\log\left[\det\left(\hat{\sigma}^2(I - H)\right)\right]. \qquad (4.15)$$

The result in (4.15) is zero under the assumption $\varepsilon \sim N(0, \sigma^2 I)$. This is shown as follows:

$C_1(\bullet)$ is undefined, since the determinant of (I-H)=0. Replacing |I-H| by $\prod_{j=1}^{n}\lambda_j$, the product of nonzero eigenvalues, we have that $\log|I-H| = \log\prod_{j=1}^{n}\lambda_j = \log(1)=0$. Therefore,

$$\frac{n}{2}\log\left[\frac{\sum\lambda}{n}\right] - \frac{1}{2}\log(\prod\lambda) = 0, \qquad (4.16)$$

and $C_1(\hat{\Sigma}(\hat{\varepsilon})) = 0$. Substituting the results (4.10) and (4.16) into (4.8),

we have:

$$\text{ICOMP}(\text{GMDH}_{\text{normal}}) = n\log(2\pi) + n\log(\hat{\sigma}^2) + n + 2C_1((X'X)^{-1}). \quad (4.17)$$

4.2.3 Derivation of ICOMPIFIM($\text{GMDH}_{\text{normal}}$)

If we adopt the second approach to quantifying model complexity in a model evaluation criterion, ICOMPIFIM, the Fisher information matrix and its inverse must be obtained. For the model in (4.1), under the assumption of normality, we have:

$$\text{Cov}(\hat{\beta}, \hat{\sigma}) \equiv F^{-1} = \begin{bmatrix} \sigma^2 (X'X)^{-1} & 0 \\ 0' & \dfrac{2\sigma^4}{n} \end{bmatrix} \quad (4.18)$$

From the definition of ICOMPIFIM in (3.11), we can obtain ICOMPIFIM for the GMDH algorithm as follows:

$$\text{ICOMPIFIM}(\text{GMDH}_{\text{normal}}) = -2\log L(\hat{\beta}, \hat{\sigma}^2) + 2C_1(\hat{F}^{-1})$$

$$= n\log(2\pi) + n\log(\hat{\sigma}^2) + n + (p+1)\log\left[\frac{\text{tr}[\hat{\sigma}^2(X'X)^{-1}] + 2\hat{\sigma}^4/n}{p+1}\right] \qquad (4.19)$$

$$-\log\left|\hat{\sigma}^2(X'X)^{-1}\right| - \log\frac{2\hat{\sigma}^4}{n}.$$

Since $C_1(\bullet)$ is scale invariant, we can factor $\hat{\sigma}^2$ from $C_1(\hat{F}^{-1})$ and simplify to obtain:

$$\text{ICOMPIFIM}(\text{GMDH}_{\text{normal}}) = n\log(2\pi) + n\log(\hat{\sigma}^2) + n$$

$$+(p+1)\log\left[\frac{\text{tr}(X'X)^{-1} + 2\hat{\sigma}^2/n}{p+1}\right] - \log\left|(X'X)^{-1}\right| - \log\frac{2\hat{\sigma}^2}{n}. \qquad (4.20)$$

From (4.18) it is easily seen that as the number of parameters, p, increases (i.e., as the size of the matrix of independent variables increases), the error variance, $\hat{\sigma}^2$, gets smaller even though the complexity gets larger. Also, as $\hat{\sigma}^2$ increases, $(X'X)^{-1}$ decreases.

Therefore, $C_1(\hat{F}^{-1})$ achieves a trade-off between these two extremes and guards against multicollinearity.

4.3 Consistency of ICOMP and ICOMPIFIM for the GMDH Algorithm

Following the work of Bozdogan and Haughton (1998), let Q represent a positive definite matrix. Under the assumption of normality, it is reasonable to assume (see Nishii, 1984, and Stewart, 1991) that, for the model in (4.1), $(X' X) / n \rightarrow Q$ as $n \rightarrow \infty$. Thus, if we have two competing models, m_{k1} and m_{k2}, where

$$m_{k1} \equiv \{\beta_0, \beta_1, \ldots, \beta_{k1}, \sigma_1^2\},$$

$$m_{k2} \equiv \{\beta_0, \beta_1, \ldots, \beta_{k2}, \sigma_2^2\}$$

and k2>k1, then we have $(X_{k1}' X_{k1})/n \rightarrow Q_1$ and $(X_{k2}' X_{k2})/n \rightarrow Q_2$, where Q_1 and Q_2 are positive definite matrices. Using this information, we show that, if m_{k2} contains the true values of the

parameter vector and m_{k1} does not, then the probabilities that both ICOMP($GMDH_{normal}$) and ICOMPIFIM($GMDH_{normal}$) select m_{k2} over m_{k1} converge to one as $n \to \infty$.

Proposition: Let m_{k1} and m_{k2} be two competing models such that the true value of the parameter vector lies in m_{k2}, but not in m_{k1}. Assume that the matrices $(X_{k1}' X_{k1})/n \to Q_1$ and $(X_{k2}' X_{k2})/n \to Q_2$, where Q_1 and Q_2 are positive definite. The probabilities that ICOMP($GMDH_{normal}$) and ICOMPIFIM($GMDH_{normal}$) select m_{k2} converge to one as $n \to \infty$.

Proof: *Part 1. ICOMP($GMDH_{normal}$).* Let $\text{maxlog}L_1$ and $\text{maxlog}L_2$ be the maximum log likelihoods on m_{k1} and m_{k2}, respectively. Also, let $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ be the maximum likelihood estimators for σ^2 (the true parameter value) on m_{k1} and m_{k2}, respectively. By Haughton (1991),

there exists a positive value, α, such that as $n \to \infty$,

$\log(\hat{\sigma}_1^2) > \alpha + \log(\hat{\sigma}_2^2)$ with probability one. Thus,

$$\text{maxlog}L_1\text{-maxlog}L_2=[n\log(2\pi) + n\log(\hat{\sigma}_1^2) + n] - [n\log(2\pi)+ n\log(\hat{\sigma}_2^2) + n]$$

$$=n[\log(\hat{\sigma}_1^2) - \log(\hat{\sigma}_2^2)] > n\alpha \qquad (4.21)$$

with probability one as $n \to \infty$. Now, since C_1 is invariant under scalar multiplication, $(X_{k1}' X_{k1})/n \to Q_1$ and $(X_{k2}' X_{k2})/n \to Q_2$, then we have that $C_1[(X_{k1}' X_{k1})^{-1} / n] \to C_1(Q_1^{-1})$ and $C_1[(X_{k2}' X_{k2})^{-1} / n] \to C_1(Q_2^{-1})$ as $n \to \infty$. By the fact that $\hat{\sigma}_2^2 \to \sigma^2$ as $n \to \infty$ and $\hat{\sigma}_1^2$ does not, there exists a value λ such that as $n \to \infty$ $C_1(X_{k1}' X_{k1}) - C_1(X_{k2}' X_{k2}) > \lambda$ with probability one (Bozdogan, Haughton, 1996). ❑

Part 2. ICOMPIFIM($GMDH_{normal}$). The first part of ICOMPIFIM($GMDH_{normal}$) is the same as the first component of ICOMP($GMDH_{normal}$). Thus, (4.21) of part 1 of proof also holds for

ICOMPIFIM($GMDH_{normal}$). Now, let $\hat{F}_1^{-1}$ and $\hat{F}_2^{-1}$ represent the estimated inverse Fisher information matrices for m_{k1} and m_{k2}, respectively. From Bozdogan and Haughton (1996), there exists a constant, δ, such that, with probability one as n goes to infinity, $C_1(\hat{F}_1^{-1}) - C_1(\hat{F}_2^{-1}) > \delta$. Thus, for the GMDH algorithm under the assumption of normality,

$$\text{ICOMPIFIM}(m_{k1}) - \text{ICOMPIFIM}(m_{k2}) > n\alpha + \delta. \qquad (4.22)$$

If δ is positive, then the proof is complete. If δ is negative, then we can multiply ICOMPIFIM($GMDH_{normal}$) through by log(n)/n, since n is constant across competing models within the algorithm. With this change, it follows that $\text{ICOMPIFIM}(m_{k1}) - \text{ICOMPIFIM}(m_{k2}) > n\alpha$ with probabilities going to one as $n \to \infty$.

Therefore, with probabilities one as $n \to \infty$, ICOMP($GMDH_{normal}$) and ICOMPIFIM($GMDH_{normal}$) will choose m_{k2} over m_{k1}. ■

4.4 The GMDH Algorithm Using Information-Based Criteria: Numerical Examples

The basic structure of the original GMDH remains the same with the incorporation of information-based model selection criteria. The basic differences in the algorithm is based on the elimination of the subdivision of the data for cross-validation and in the selection of models to be passed on to the next generation. In step 2 of the algorithm, the screening stage, information-based model evaluation criteria are calculated for each model as "fitness" values. After calculating the AIC, ICOMP, or ICOMPIFIM value for each of the models generated at a level, we utilize a selection ratio to determine which models to keep for the next level of the algorithm. This selection ratio is based on the one used by Luh, Minesky, and Bozdogan (1997) in their work on genetic algorithms.

Let Max(crit)_j represent the highest value of the computed information-theoretic criterion at level j of the algorithm. Let crit_{ij} represent the criterion value for model i of level j. For each model obtained at level j, we obtain the difference:

$$Diff_{ij}=Max(crit)_j-crit_{ij}.$$

The average of these differences, $Avgdiff_j$ is obtained. The selection ratio for model i in level j is then obtained via:

$$\text{Selection Ratio}_{ij} = \frac{Diff_{ij}}{Avgdiff_j}. \qquad (4.23)$$

ICOMP and ICOMPIFIM are much more general than AIC, as previously discussed. They control the risks of both insufficient and overparameterized models and incorporate the assumption of dependence and the independence of the residuals in one criterion function. ICOMP and ICOMPIFIM relieve the researcher of any need to consider the parameter dimension of a model explicitly. This is illustrated through the numerical example provided in Section 4.3.1. Also, as discussed in Section 4.2, ICOMPIFIM inherently guards against multicollinearity in the set of data. A small numerical example is provided in Section 4.3.2 which effectively illustrates this tendency.

4.4.1 Numerical Example: Study of Vitamin B_2 in Turnip Greens

We consider the problem of identifying the appropriate model for the data set provided in Table 1.1 of Chapter 1. These data are from a study performed by Wakely (1949) on the quantity of vitamin B_2 in turnip greens. The data set consists of n=27 observation and p=3 predictor or independent variables.

Using AIC obtained in (4.7), ICOMP for the GMDH algorithm in (4.17), and ICOMPIFIM in (4.20), we determine the "best" model via the GMDH algorithm. All the computations are carried out using the newly developed open architecture algorithm and functions which is provided in Appendix B. These algorithms and functions use MATLAB® as the basis kernel. The "best" models are determined at each of three generations of the GMDH algorithm; hence, the "best" second-degree polynomial, fourth-degree polynomial, and sixth-degree polynomial forms are identified using the information-theoretic criteria. Also, the "best" global model is identified. The results are provided in Table 4.1 for three generations of the algorithm. It should be noted that the third generation signifies a

stopping point for the algorithm, since the values for all of the criteria (specifically, the value of AIC) increase in the fourth generation.

From Table 4.1, we see that the "best" model for the data based on the AIC criterion is the level 3 model:

$$\hat{y} = 44.3126 + 0.7486\bar{z}_1 - 0.6155\bar{z}_3 + 0.0025\bar{z}_1^2 + .0145\bar{z}_3^2 - .0129\bar{z}_1\bar{z}_3 \quad (4.24)$$

where $\bar{z}_1$ and $\bar{z}_3$ are variables as constructed in level two of the algorithm. As seen in Table 4.1, these variables are obtained by evaluation of the polynomials constructed in level one and, then, by evaluation of the polynomials constructed in level two. We can obtain the complex form of (4.24) in terms of the original variables (x_1, x_2, x_3) by backtracing using the backtracing function written in MATLAB®. (The backtracing function is also provided in Appendix B.)

TABLE 4.1. Model Selection for the Turnip Green Data Under the GMDH Algorithm

Variables	Level 1: Equation	AIC	ICOMP	ICOMPIFIM
(x_1,x_2)	$z_1=104.773-24.89x_1+390.783x_2+5.2x_1^2-987.841x_2^2+21.008x_1x_2$	181.9	200.1*	287.5*
(x_1,x_3)	$z_2=850.031-5.863x_1-206.188x_3+18.535x_1^2+14.48x_3^2-7.258x_1x_3$	232.8	259.9	377.1
(x_2,x_3)	$z_3=256.593+386.385x_2-39.562x_3-1205.341x_2^2+2.037x_3^2+20.069x_2x_3$	177.8	200.1*	290.6
	Level 2:			
(z_1,z_2)	$\bar{z}_1=-99.4685+5.8914z_1-1.1231z_2-0.0363z_1^2+0.0014z_2^2+0.0066z_1z_2$	226.1	314.3	491.2
(z_1,z_3)	$\bar{z}_2=-52.1378+0.4995z_1+1.5894z_3-0.003z_1^2-0.0053z_3^2+0.0025z_1z_3$	177.2	257.6	406.1
(z_2,z_3)	$\bar{z}_3=8.2319+0.0211z_2+0.7383z_3+0.005z_2^2+0.0041z_3^2-0.0024z_2z_3$	176.7	248.1	380.1
	Level 3:			
$(\bar{z}_1,\bar{z}_2)$	$\bar{\bar{z}}_1=5.4847+0.0664\bar{z}_1+1.221\bar{z}_2+0.0011\bar{z}_1^2-0.0032\bar{z}_2^2-0.0011\bar{z}_1\bar{z}_2$	195.7	NA**	NA**
$(\bar{z}_1,\bar{z}_3)$	$\bar{\bar{z}}_2=44.3126+0.7486\bar{z}_1-0.6155\bar{z}_3+0.0025\bar{z}_1^2+0.0145\bar{z}_3^2-0.0129\bar{z}_1\bar{z}_3$	175.5	NA**	NA**
$(\bar{z}_2,\bar{z}_3)$	$\bar{\bar{z}}_3=16.259+0.4656\bar{z}_2+0.0601\bar{z}_3+0.0029\bar{z}_2^2+0.0123\bar{z}_3^2-0.012\bar{z}_2\bar{z}_3$	176.5	NA**	NA**

Notes: *Global minima of the criteria

**Level 2 is the stopping level for ICOMP and ICOMPIFIM; hence, the algorithm does not proceed to Level 3 for these criteria.

With ICOMP and ICOMPIFIM, the "best" models are is identified to be either a second-degree polynomial in x_2 and x_3 or a second-degree polynomial in x_1 and x_2:

ICOMP:

$$\hat{y} = 256.593 + 386.385x_2 - 39.562x_3 - 1205341x_2^2 + 2.037x_3^2 + 20.069x_2x_3 \quad (4.25)$$

ICOMP and ICOMPIFIM:

$$\hat{y} = 104.773 - 24.89x_1 + 390.783x_2 + 5.2x_1^2 - 987.841x_2^2 + 21.008x_1x_2 \quad (4.26)$$

In this example, ICOMP and ICOMPIFIM choose a simpler model (4.25 or 4.26) to approximate the data. The model identified by AIC (4.24) is a much more complex, less parsimonious model than those chosen by ICOMP or ICOMPIFIM. As noted in the derivation of AIC for the GMDH algorithm in Section 4.2.1, the penalty function (i.e., 2*number of parameters) remains constant at each level of the algorithm even though the true number of parameters in the model increases with each generation. Also, at each level of the algorithm, the profusion of complexity (i.e., interdependence among the parameter estimates) increases with each generation. Therefore, as

seen through observing the results in Table 4.1, AIC does not appropriately distinguish model complexity and lack of parsimony in the GMDH environment.

The ICOMP and ICOMPIFIM criteria both incorporate a measure of the complexity due to the covariance of the parameter estimates and the covariance of the error terms within the environment of the GMDH algorithm. While under the normal error assumption on the random error, the complexity of the covariance matrix of the errors will be zero, as seen in (4.16). However, when the errors are autocorrelated or when the errors have other general covariance structures, this term (4.16) will not be zero. Given these results, we see that the criteria which incorporate model complexity into the penalty term are preferable for use within the GMDH algorithm.

This same set of data was used by Anderson and Bancroft (1959) to fit the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \varepsilon.$$

Fitting this model we obtain the estimated regression model:

$$\hat{y} = 82.44 + 2.62x_1 - 75.69x_2 + 1.52x_3 - 1.64x_1x_2 \text{ with } R^2 = 0.754. \quad (4.27)$$

Later, Draper and Smith (1981) analyzed the same data using a stepwise regression technique. After correcting their error in analysis, the regression polynomial obtained was:

$$\hat{y} = 121.1 + 496.18x_2 - 5.79x_3 - 1129x_2^2 \text{ with } R^2 = 0.903. \quad (4.28)$$

The model chosen by ICOMP and ICOMPIFIM (4.26), without using any arbitrary assumption of model form prior to data modeling modeling, has R^2=0.9201. Thus, the GMDH with information-theoretic model selection criteria provides the analyst with an objective way to determine the functional form of their model directly from the data in a data-adaptive fashion, as opposed to relying on the many locally sub-optimal model search methods.

4.4.2 Numerical Example: Study of Demand of Laundry Detergent

We consider the problem of identifying the appropriate model for the data set provided in Table 1 of Appendix A. These data are from a study on the demand of a brand of liquid laundry detergent. (Bowerman et al., 1986) The data set consists of n=30 observations and p=4 predictor or independent variables. The dependent variable is:

y = demand (in hundreds of thousands of bottles) in a four-week sales period.

The independent variables are:

x_1 = price (in dollars) of a bottle of the detergent;

x_2 = average industry price (in dollars) of competitive brands;

x_3 = advertising expenditure (in hundreds of thousands of dollars) in the sales period;

x_4 = price difference between detergent's price in sales period and average industry price during the same period (i.e., $x_4 = x_2 - x_1$).

In level one of the algorithm, six polynomials are constructed. These six polynomials and the associated computed values of AIC, ICOMP, and ICOMPIFIM are provided in Table 4.2. In the screening stage of level one, the values of the selection ratio for each of the six models are:

New Variable	Selection Ratio (AIC)	Selection Ratio (ICOMP)	Selection Ratio (ICOMPIFIM)
$z_1=f(x_1,x_2)$	.0902	.7184	.756
$z_2=f(x_1,x_3)$	0	0	0
$z_3=f(x_1,x_4)$	.0902	.7129	.749
$z_4=f(x_2,x_3)$	2.044**	1.502**	1.50**
$z_5=f(x_2,x_4)$	.0902	.6570	.685
$z_6=f(x_3,x_4)$	3.685**	2.409**	2.309**

** Those models with a selection ratio value greater than one (i.e., those models to be passed to the second generation)

As seen in the above summary, two models are identified for passing on to the second generation of the algorithm. These two models are z_4 and z_6. Thus, in level two of the algorithm, only one polynomial is constructed as a function of z_4 and z_6:

For both ICOMP and ICOMPIFIM, the minimum value of the criteria occurs at z_6 of level one. Thus, the "best" model is:

$$z_6 = 32.098 - 8.637x_3 + 14.744x_4 + 0.7594x_3^2 + 1.1074x_4^2 - 2.104x_3x_4, \quad (4.29)$$

with an R^2 of .9226. For AIC, the criterion value in generation 2 for the model (4.29) is -3.857. Thus, AIC chooses as the "best"

TABLE 4.2. Model Selection in Level One for the Detergent Data Under the GMDH Algorithm with Information-Theoretic Criteria

Variables	*Equation*	*AIC*	*ICOMP*	*ICOMPIFIM*
(x_1,x_2)	$z_1=36.14+4.27757x_1-16.972x_2-4.64x_1^2-0.703x_2^2+6.714x_1x_2$	16.6	52.0	92.5
(x_1,x_3)	$z_2=-123.74+68.046x_1+0.2767x_3-7.868x_1^2+0.4985x_3^2-1.449x_1x_3$	17.11	65.07	120.2
(x_1,x_4)	$z_3=36.14-12.6966x_1-16.972x_4+1.372x_1^2-0.703x_4^2+5.308x_1x_4$	16.6	52.1	92.74
(x_2,x_3)	$z_4=-5.477+9.801x_2-3.45x_3-0.164x_2^2+0.698x_3^2-1.146x_2x_3$	5.574	37.74	65.27
(x_2,x_4)	$z_5=36.14-12.697x_2-4.276x_4+1.372x_2^2-4.639x_4^2+2.563x_2x_4$	16.6	53.11	95.1
(x_3,x_4)	$z_6=32.098-8.637x_3+14.744x_4+0.759x_3^2+1.107x_4^2-2.104x_3x_4$	3.684**	21.24**	35.69**

**The "best" model in level one as identified from the GMDH algorithm.

approximating model a quadratic polynomial containing x_2, x_3, and x_4 with an R^2 of .923. The exact form of this model after backtracing is:

$$\begin{aligned}\bar{z}_1 = &-1447.8955 + 784.8607x_2 + 390.0911x_3 - 1137.7381x_4 - 116.9148x_2^2 \\ &-29.3251x_3^2 - 309.5824x_4^2 + 3.4799x_2^3 + 0.4045x_3^3 - 33.6689x_4^3 \\ &-0.0292x_2^4 + 0.002x_3^4 - 1.2645x_4^4 - 198.0332x_2x_3 + 306.1178x_2x_4 \\ &+317.309x_3x_4 + 13.4122x_2x_3^2 + 22.9931x_2x_4^2 + 75.5878x_3x_4^2 \\ &+26.0482x_2^2x_3 - 5.1336x_2^2x_4 - 23.3702x_3^2x_4 - 1.4355x_2^2x_3^2 \\ &-0.3856x_2^2x_4^2 - 4.6599x_3^2x_4^2 - 0.1154x_2x_3^3 + 4.8043x_3x_4^3 \\ &-0.4069x_2^3x_3 + 0.1832x_3^3x_4 - 79.4793x_2x_3x_4 + 0.7325x_2^2x_3x_4 \\ &+5.1082x_2x_3^2x_4 - 2.6889x_2x_3x_4^2. \end{aligned} \tag{4.30}$$

Obviously, the two criteria, ICOMP and ICOMPIFIM, choose a much more parsimonious model.

4.5 Discussion of the GMDH with Information Theoretic Criteria

The GMDH algorithm with information-theoretic model evaluation criteria provides a stronger, more robust version of the original algorithm. The two greatest benefits of the use of the

information-theoretic criteria are the elimination of the need for subdivision of the data and the selection of models based on goodness of fit as well as on model complexity. In Chapter 2, the Wakely turnip data was modeled using the original GMDH with the regularity criterion as the selection criterion. The first subdivision of the data in Table 2.1 yielded

$$y_3 = 369.85 + 476.74x_2 - 67.41x_3 - 1302.06x_2^2 + 3.7x_3^2 + 13.94x_2x_3$$

as the best approximating model. The second subdivision of data yielded

$$y_1 = 108.25 - 29.22x_1 + 284.87x_2 + 6.56x_1^2 - 812.22x_2^2 + 36.23x_1x_2$$

as the best approximating model.

We note in the previous section that the GMDH with ICOMP chooses both a model in x_1 and x_2 (4.25) as well as one in x_2 and x_3 (4.26) as good approximating models. ICOMPIFIM chooses the model in x_2 and x_3 (4.26) as the best fitting model. Thus, we obtained very

similar results without any subdivision of the data when using ICOMP or ICOMPIFIM.

We also note that AIC does not take into account the complexity of the chosen model. Hence, it chooses much more complex, less parsimonious models as seen in both examples in Section 4.4. Thus, in order to resolve the issue of overspecification of models that was identified with respect to the original GMDH, either ICOMP or ICOMPIFIM should be used within the framework of the GMDH algorithm.

As can be seen by the model chosen in (4.30) by AIC for the detergent data, the GMDH produces highly complex models as it moves up the hierarchy of generations. In fact, one of the arguments against the use of the GMDH is that it results in overparameterized models. There are several proposed versions of GMDH which attempt to deal with the issue of overparameterization by changing some of the steps or by changing the basic form of the quadratic polynomial to different types of functions. Readers are referred to Ikeda et al. (1976) and Tamuru and Kondo (1984) for some of these variations. As opposed to changing the structural steps of the algorithm or the

functional form of the polynomial, we propose the use of the genetic algorithm as a means to eliminate those terms within a chosen model which are not necessary to effectively using a model for predictive reasons. In the next chapter, we present an overview of the genetic algorithm and discuss its usefulness within the framework of the GMDH algorithm.

CHAPTER 5

THE GENETIC ALGORITHM WITHIN THE FRAMEWORK OF THE GMDH ALGORITHM

At the end of the GMDH algorithm, one or more "best" approximating models are identified. This "best" approximating model is of the Kolmogorov-Gabor form shown in (2.1). Given the pairwise nature of the construction of models at each level, a data set with a large number of potential predictor variables may frequently result in a chosen model at the third, fourth, or higher generation of the algorithm. After backtracing, if we started with six input variables and a model is chosen at the third generation of the algorithm, then we would have an Ivakhnenko polynomial of degree eight in $x_1, x_2, ..., x_6$. Thus, the resulting polynomial would have 3,003 terms with hundreds of terms like $\beta_k x_1^2 x_2^3 x_3^5 x_4^8 x_5^6 x_6^3$ and $\beta_l x_2^4 x_4^4$. Therefore, a great concern about the algorithm is the

number of potentially unnecessary terms that are maintained in the "best" approximating model at the end of the heuristic search in the hierarchy.

Much work has been done to attempt to reduce the number of terms that are unnecessary for maintaining the predictive power of the resulting model. Barron et al. (1984) used a clustering algorithm to identify subsets of "like" predictor variables. They also proposed that the original candidate inputs ($x_1,...,x_m$) be introduced as input variables into each following generation along with the outputs from the previous level's generation of new models. The main weakness of this approach is the number of additional models that must be estimated at each level of the algorithm.

Tamura and Kondo (1984) proposed the identification of both partial and intermediate polynomials at each generation to allow for some choice in the terms included in the following layer of the algorithm. At each level, the optimal partial polynomials are determined by applying a stepwise regression procedure to each second-order Ivakhnenko polynomial generated. However, there exists much criticism in the literature (Moses, 1986, and Wilkinson,

1989) of stepwise selection since it does not always find the best subset of predictor variables from the original set of variables. Also, stepwise regression analysis only uses local sampling in its searching process (Sokal and Rohlf, 1981). In order to be able to search the large solution space and to keep from introducing additional variables at each generation of the algorithm, we propose the use of the genetic algorithm with information-based criteria to determine the "best" subset of the variables (i.e., terms) included in the final model obtained at the end of the algorithm.

5.1 An Overview of the Genetic Algorithm

The genetic algorithm (GA) is a general purpose search algorithm that is loosely based on the mechanisms of biological inheritance and how traits are transmitted from the parents to their offspring. Genetic algorithms were formally introduced in the United States in the 1970s by John Holland at The University of Michigan. Thus, much of the early work on the GA was done by Holland (1975) who proposed that the biological methods of crossover, mutation, and

reproduction could be used as a computer-based search and problem solving method. The following discussion of the GA is merely a brief overview of the basic nature of the algorithm. Readers who desire a more in-depth discussion of GAs are referred to Goldberg's (1989) book Genetic Algorithm in Search, Optimization, and Machine Learning and Holland's (1992) article *Genetic Algorithms.*

GA's belong to the family of stochastic search methods. The major difference between the GA and other stochastic search methods is the operation of the method on a population of solutions at each iteration as opposed to operating only on a single solution.

Following the language of biological evolution, each potential solution to the problem is encoded as a chromosome. An initial population of chromosomes is created. Each chromosome represents a complete solution to the objective function or the problem being solved. The algorithm then applies genetic operators--crossover and mutation--to the individuals in the population to generate new chromosomes (i.e., offspring or children) from pieces of the parent chromosomes. The crossover operator combines two parent chromosomes to produce two more chromosomes. Thus, it is through

the crossover function that genetic material is passed from one generation to the next. Through the mutation operator, randomness is introduced into the search. This randomness ensures that additional solutions are created and evaluated that might not be found by crossover alone. (Goldberg, 1989)

In general, a GA consists of a set of basic, sequential steps (Lin and Lee, 1996, and Bearse and Bozdogan, 1998):

1. Creation of an initial population of chromosomes.
2. Evaluation of each of the chromosomes in the population.
3. Creation of offspring by application of the genetic operators mutation and crossover.
4. Elimination of the unfit members of the population.
5. Evaluation of the new offspring.
6. Insertion of offspring into the population.
7. Evaluate stopping criterion. If satisfied, stop. Otherwise, return to step 3.

The GA consists of three major components: the representation of the problem as a string of digits, the fitness function, and the genetic operators. The effectiveness and efficiency of the GA is

dependent upon the development of these three components. The overall performance of the GA is highly dependent upon the development of the fitness function, since it is this function which evaluates the quality of a potential solution and determines a potential solution's likelihood of reproducing. It is through this fitness function that the biological idea of "survival of the fittest" is defined so that the "fittest" potential solutions are kept from one generation to the next and the least "fit" potential solutions die off. In many cases, the GA performs very efficiently. In fact, since the GA does not use gradients, it often performs better than gradient search methods, especially if the search space has a lot of local optima. (Goldberg, 1989) As stated best by Nygard, Ficek, and Sharda (1992), "...the efficiency of genetic search is due to an 'intrinsic parallelism'--each trial becomes less and less random, as desirable substrings increase in frequency of occurrence and undesirable substrings are suppressed."

5.2 Genetic Algorithms for the GMDH

As previously stated, the strength of the GMDH algorithm is its ability to produce a self-organized functional representation of the form $y=f(x_1,x_2,...,x_m)$ without any apriori assumptions regarding the functional form prior to the modeling of the data. Along with this strength comes one of the biggest weaknesses of the GMDH. That is, the resulting model is often quite complex with a number of higher-order terms that do not substantially improve the predictive power of the model. Thus, in this thesis, we propose and develop for the first time the use of a genetic algorithm for choosing the best terms (i.e., predictors) to maintain in the final model. Since the GA is a stochastic search process, it allows us to globally search for the best "sub-model" even when the initial set of terms is quite large. There are two possible applications of a GA within the framework of the GMDH. The first is the use of the GA only at the end of the GMDH to identify the "best" sub-model of the final version produced by the algorithm. The second application is to use the GA prior to invoking the GMDH in order to reduce the number of variables introduced into

the first level of the GMDH. Both applications are presented and the results are compared in Chapter 6 of this thesis.

5.2.1 A Genetic Algorithm Using Information-Theoretic Fitness Functions for the Backtraced GMDH Model

The basic steps of the described algorithm follows the approach of Goldberg (1989), often referred to as the Simple Genetic Algorithm (SGA). The more detailed components of the approach, including the incorporation of ICOMP and ICOMP(IFIM) for the fitness functions within the GA, follow the work of Luh, Minesky, and Bozdogan (1996). The major difference in the approaches lies in the initialization of the population. In the use of the GA for multiple regression, the initial population consists of various combinations of predictors in different possible functional forms of a regression equation. However, in our application, we are provided with the Kolmogorov-Gabor functional form (2.1) as an output from backtracing the model obtained via the GMDH algorithm. Thus, the possible models to be evaluated are limited to subsets of the model generated by the GMDH.

The first stage of the GA is to establish the coding scheme for the possible sub-models of the resulting GMDH equation. For example, let us assume that the "best" approximating model from the GMDH was a second-generation Kolmogorov-Gabor equation in three variables. Then, the structure of the model would be:

$$y = a_i + \sum_{i=1}^{3} b_i x_i + \sum_{i=1}^{3} c_i x_i^2 + \sum_{i=1}^{3} d_i x_i^3 + \sum_{i=1}^{3} f_i x_i^4 + \sum_{i=1}^{3}\sum_{i\neq j=1}^{3} g_i x_i x_j + \sum_{i=1}^{3}\sum_{i\neq j=1}^{3} h_i x_i^2 x_j$$

$$+ \sum_{i=1}^{3}\sum_{i\neq j=1}^{3} l_i x_i^2 x_j^2 + \sum_{i=1}^{3}\sum_{i\neq j=1}^{3} m_i x_i^3 x_j + \sum_{i=1}^{3}\sum_{i\neq j=1}^{3}\sum_{j\neq k=1}^{3} n_i x_i x_j^2 x_k \qquad (5.1)$$

The estimated model (5.1) contains 33 terms, including the intercept term. Thus, each chromosome is represented by a string of 33 binary digits. Each possible sub-model is represented as a binary string, where (1) indicates the presence of a term in the model and (0) indicates the absence of that term. So, the binary string {111111111111111111111111111111111} represents the case where all 33 terms are included in the model being evaluated, and {111100000000000000000000000000000} represents the case

where only the intercept term, x_1, x_2, and x_3 are included in the submodel (i.e., none of the interaction or higher-order terms are included).

The second major stage of the GA is to initialize a population of chromosomes. The size of the population (n) is a set of n submodels selected at random by the algorithm to enter into the first iteration of the GA. It should be noted that, for our computational modules of the GA, the population size is selected by the user. However, as is seen in Chapter 6, the size of the population does affect the ability of the algorithm to identify the "best" submodel within a reasonable number of generations.

For the fitness function used to evaluate each model (i.e., chromosome) generated, we use the information-theoretic criteria of ICOMP or ICOMP(IFIM). These criteria were defined in general in equations (3.10) and (3.11), respectively. The derived forms of these two criteria are those for the univariate normal multiple regression case. As derived in Bozdogan (1988) and Bozdogan and Haughton (1998), the ICOMP and ICOMPIFIM evaluation criteria used within the GA are, respectively:

$$\begin{aligned} \text{ICOMP(GA)} &= -2\log L\left(\hat{\beta},\hat{\sigma}^2\right) + 2C_1\left(\text{Cov}(\hat{\beta})\right) + 2C_1\left(\text{Cov}(\hat{\varepsilon})\right), \\ &= n\log(2\pi) + n\log(\text{RSS}/n) + n + 2C_1\left((\mathbf{X}'\mathbf{X})^{-1}\right). \end{aligned} \tag{5.2}$$

$$\begin{aligned} \text{ICOMPIFIM(GA)} &= -2\log L\left(\hat{\beta},\hat{\sigma}^2\right) + 2C_1\left(\text{Cov}(\hat{\mathbf{F}}^{-1})\right) \\ &= n\log(2\pi) + n\log(\hat{\sigma}^2) + n + (k+1)\log\left[\frac{\text{tr}[\hat{\sigma}^2(X'X)^{-1}] + 2\hat{\sigma}^4/n}{k+1}\right] \\ &\quad - \log\left|\hat{\sigma}^2(X'X)^{-1}\right| - \log\frac{2\hat{\sigma}^4}{n}. \end{aligned} \tag{5.3}$$

In the algorithm and functions we provide for the GA, the investigator chooses whether the desired fitness function is ICOMP(GA) or ICOMPIFIM(GA).

Thus, the GA evaluates the chromosomes at each generation using either ICOMP(GA) or ICOMPIFIM(GA), as chosen by the investigator. The chosen criterion is evaluated for each submodel in order to assess which of the generated submodels are "fittest" and are therefore used in the mating pool. Following Luh, Minesky, and Bozdogan (1997), a fitness ratio is computed to determine the likelihood of a given submodel being introduced into the mating pool

and, thereby, being used as parents to produce offspring. The fitness ratio for submodel i is based on the difference between the maximum value of the information criteria in the existing population and the value of ICOMP or ICOMPIFIM for submodel i. As shown in (4.23), the fitness ratio for submodel i in a population of size n is:

$$\text{Fitness Ratio}_i = \frac{\text{Diff}_i}{\sum_{i=1}^{n} \text{Diff}_i / n}, \tag{5.4}$$

where $\text{Diff}_i = \underset{\forall n}{\text{Max}}\,(\text{Criterion Value}) - (\text{Criterion Value}_i)$. The higher the fitness ratio for a submodel, the higher the likelihood that the submodel will be included in the mating pool. So, if submodel 1 has a fitness ratio of 0.5 and submodel 2 has a fitness ratio of 1.5, then submodel 2 is three times more likely to enter into the mating pool.

The last major components that must be defined for the GA are the genetic crossover and mutation operators. The crossover probability, another input by the investigator, determines the probability of mating between two parents in the mating pool. Thus, a crossover probability of zero (0) means that no mating between

parents occurs, and all of the members of the current mating pool are entered into the next generation of the algorithm. In this case, the only possibility for change in the population comes through the mutation operation. A crossover probability of one (1) indicates that all parents mate to form new offspring. Here, none of the parents are maintained and passed into the next generation. The risk with an extremely high crossover probability is that a "good" submodel is removed from the search prior to possible genetic improvements that might be made in future generations of the algorithm. In general, there are no specific rules for identifying the best crossover probability.

The basic crossover operator used in our algorithm is based on that of Goldberg (1989). As illustrated in Figure 5.1, given a pair of parent chromosomes, the algorithm randomly chooses a position along the chromosomes to be the crossover point. The pieces of the parent chromosomes to the right of the crossover point are interchanged in order to create two offspring.

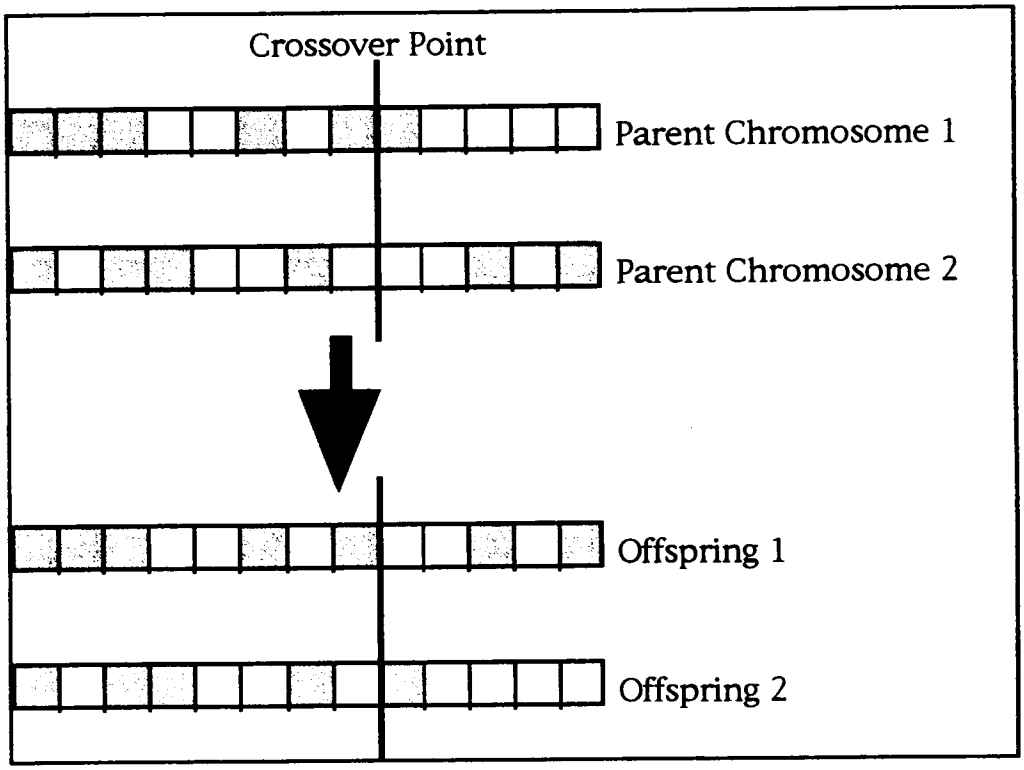

FIGURE 5.1 The Crossover Operator Used in the Genetic Algorithm

The mutation operator is used to allow for a drastic move of the GA into a completely different response surface area. The mutation probability determines the likelihood that the algorithm will randomly add or delete a term from the submodel (i.e., a randomly selected binary digit is changed from 0 to 1 or from 1 to 0). Lin and Lee (1996) recommend that the probability of mutation is kept fairly low in order to prevent the GA from becoming merely a random search technique.

After these general components are defined, the GA moves from one generation to the next for the specified number of generations. Each generation is made up of a mixture of parent submodels and offspring, as determined by the crossover and mutation ratios. Bearse and Bozdogan (1998) point out that the relationship between the population size (n), the number of alleles (i.e., binary digits) in each chromosome, and the structure of the data set influence the number of generations needed for the GA to reach an optimal solution. This relationship is currently not completely understood. In our results shown in Chapter 6, we compare the performance of the GA on the GMDH backtraced model for varying population sizes and number of generations the algorithm is run.

5.2.2 A Numerical Example: The GA with ICOMP Fitness Function for the Backtraced GMDH Model

We illustrate the application of the GA to increase the parsimony of the resulting model from the GMDH by using the Wakely (1949) data from Table 1.1. Recall that this data set contains only three predictor variables. Thus, for this problem, all possible

submodels can easily be evaluated since only 2^6-1 submodels are possible. However, we use this data set to walk through the major components of the GA in this environment. Recall from Section 4.3.1 that the GMDH with ICOMP and ICOMPIFIM identified the following model (4.25) for the Wakely (1949) data:

$$\hat{y} = 104.773 - 2.89x_1 + 390.783x_2 + 5.2x_1^2 - 987.841x_2^2 + 21.008x_1x_2.$$

Following the major steps of the GA stated in Section 5.1, we begin the algorithm with the creation of an initial population of chromosomes. Since the above model contains six terms, we encode each possible submodel as a binary string with six loci, as shown in Table 5.1.

Also in the first step is the selection of an initial population of chromosomes. For this example, we choose a population size of six (i.e., n=6). From the 63 possible submodels, the algorithm randomly chooses six submodels (i.e., chromosomes) to use in the first generation of the algorithm. These six chromosomes in binary form

TABLE 5.1 Possible Submodels and Their Binary Representation for the GA

Terms in Submodel	Binary String
Intercept, $x_1, x_2, x_1^2, x_2^2, x_1x_2$	111111
Intercept, x_1, x_2, x_1^2, x_2^2	111110
Intercept, x_1, x_2, x_1^2	111100
Intercept, x_1, x_2	111000
⋮	⋮
x_1x_2	000001

are M_1=[1000001], M_2=[111110], M_3=[100101], M_4=[001101], M_5=[010110], and M_6=[111110]. We note that M_2=M_6. In our algorithm, we do allow for duplicate members in a population. However, if desired, the investigator can require six unique chromosomes in the initial population.

The second stage is the evaluation of each of these six chromosomes in the initial population. Using ICOMP in (5.2) as the fitness function, the values of the fitness function for each of the chromosomes in generation 1 are shown in column 3 of Table 5.2. Since the creation of offspring is dependent on the selection of a

mating pool, we compute the fitness ratio (5.4) for each of the chromosomes in generation one. These values of the fitness ratio are shown in column 4 of Table 5.2.

The third step of the GA is the creation of the offspring through the crossover and mutation genetic operators. The first pair of parent chromosomes chosen to mate are {M_3 and M_5}; the second set of parents are {M_2 and M_6}; and the third pair of mating parents are

TABLE 5.2. Generations One and Two Chromosomes and their Fitness Values for the GMDH Model of the Wakely Data

Generation 1 Model ID	Binary String	ICOMP	Fitness Ratio
M_1	100001	207.02	1.25
M_2	111110	199.98	1.36
M_3	100101	202.73	1.32
M_4	001101	283.83	0
M_5	010110	240.69	0.70
M_6	111110	199.98	1.36
Generation 2			
M_1	100110	198.61	1.46
M_2	010101	235.50	0.18
M_3	111110	199.98	1.41
M_4	111010	196.13	1.54
M_5	111110	199.98	1.41
M_6	010110	240.69	0

{M_2 and M_5}. For each pair of parents, the crossover point (i.e., the point in the binary string where the crossover occurs) is chosen randomly by the computer algorithm. As shown in Figure 5.1, a new population of submodels is formed by interchanging the chromosome alleles to the right of the crossover point. After the crossover operator occurs, each allele in the new chromosomes has a possibility of mutation, based on the mutation probability provided by the investigator. (In this example, the mutation probability is set to be 0.001.) If, however, both parents have the identical binary string, then we increase the mutation probability to 0.10 to prevent the algorithm from getting stuck in one particular region of the overall solution space. The results for the mating in generations one and two are provided in Table 5.3.

Those members of the population which do not get selected to mate get eliminated from the population. Hence, each new generation after generation one begins with the offspring from the previous generation. The algorithm continues until the specified number of generations has been completed or until the stopping criterion has been met.

TABLE 5.3. Mating Results for Generations One and Two of the GA for the Wakely Data Model

Generation 1 Parents	Crossover Pt.	Offspring	Generation 2 Parents	Crossover Pt.	Offspring
1001|01	4	100110	1|11110	1	111110
0101|10		010101	1|(1)1110		101110
11111|0	5	111110	|111110	0	111110
111(1)1|0*		111010	|111010		111010
111|110	3	111110	10011|0	5	100110
010|110		010110	10011| (0)		100111

*Note: (1) denotes mutation of the allele from a 1 to a 0. (0) denotes mutation from a 0 to a 1. So, in the second set of parents in generation 1, the second mate mutates at the fourth allele.

5.3 Using the GA with Information-Theoretic Criteria to Pre-Screen Variables Prior to the GMDH

For a data set with a large number of potential predictor variables, the GMDH tends to move to a higher number of generations in order to incorporate the variables into the model. This problem of selection of appropriate subsets of predictor variables prior to the modeling of the data is a long outstanding issue. It has been at the root of many studies done in the work on

regression analysis (Mantel, 1970, Boyce et al., 1974, Hocking, 1976, and Miller, 1990). All possible subset routines are often used to identify the subset of predictor variables. However, when a data set consists of 20 potential predictor variables, the number of possible subsets is 1,233,311. Therefore, it is both practically and computationally extremely inefficient to attempt to evaluate all possible subsets prior to the modeling of the data.

A second common approach for variable selection is to use stepwise regression routines. Again, much criticism of stepwise algorithms exists in the literature. As with many subset selection routines, stepwise regression does not necessarily find the optimum subset of predictor variables (Hocking, 1976 and 1983, Moses, 1986). A key weakness lies in the nature of the stepwise regression algorithms, which primarily employ local searching. Hence, stepwise routines often converge at a "local optimum" within a limited area of the solution space. (Sokal and Rohlf, 1981)

We propose the use of the GA with information-based model evaluation criteria as an effective means for identifying the subset of variables to be entered into the GMDH algorithm. As previously

stated, the GA allows us to perform a stochastic search which performs well in the face of a large solution space with potential local optima (assuming a sufficient population size and number of generations).

The basic structure of the proposed GA is the same as that presented in the previous section. The major difference lies in the nature of the input into the algorithm. When the GA was utilized to identify a more parsimonious submodel to the one obtained from the GMDH, the functional form of the response surface was known prior to invoking the GA. In this case, the nature of the functional form is unknown prior to variable subset selection. Therefore, we assume a functional form linear in the predictor variables entered into the GA. For example, if the potential set of predictor variables is $X=[x_1,x_2,x_3,x_4,x_5]$, then the submodels being evaluated by the GA are subsets of the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$. We utilize the GA, guided by information-based evaluation criteria, to help select the best variables to enter in the GMDH algorithm. As is shown in the following chapters, this approach tends to reduce the computational burden faced by the GMDH.

To illustrate the use of the GA for variable screening, we use a data set containing seven potential predictor variables. The data set, originally from G. Thackray of ICI plc, is used by Wetherill et al. (1986) to fit a response surface model. The data set, provided in Appendix A, is from a study on the concentration of gas output from a liquid stream. The response variable, y, is the concentration of gas output, ppm. The seven potential predictor variables are:

x_1=vessel pressure
x_2=nozzle pressure
x_3=flow rate
x_4=inlet gas concentration (ppm)
x_5=liquid temperature (C^o)
x_6=pH
x_7=liquid concentration.

Without pre-screening of the variables, the GMDH algorithm moves to the fifth generation and provides a model of the tenth-order. In order to identify the "best" subset of predictor variables prior to invoking the GMDH, we enter the variables $x_1, x_2, ..., x_7$ into the GA. (Note: Due to the scale of x_4, we take the natural log of x_4 prior to entering it into the GA.) We run the GA for 30 generations

with a population size of n=8. The algorithm converges on $\{x_1, x_3, x_4,$ and $x_7\}$ as the subset of variables with the minimum value of ICOMP being 585.6. Using these variables as the four predictor variables entered into the GMDH with ICOMP as the model evaluation criteria, the GMDH now converges on a Level 3 model in four variables. Still, the resulting model from the GMDH is an eighth-order model containing 294 terms. The resulting value of ICOMP for this GMDH model is 620.636, obviously higher than the criterion value identified in the subset selection GA due to the greater complexity of the model.

In order to obtain a "good" submodel with fewer terms, we again invoke the GA with information criteria, as in Section 5.2.2. The final model chosen by the GA after 200 generations with a population size of n=60 was the following with an ICOMP value of 567.9.

$$\hat{y} = b_0 + b_1x_1 + b_2x_3 + b_3x_4 + b_4x_7 + b_5x_7^2 + b_6x_1x_4 + b_7x_1x_7^2 + b_8x_1^2x_4 + b_9x_1x_3^2. \tag{5.5}$$

As is seen through this numerical example, a model which includes all terms, regardless of the contribution of the individual terms, is rarely a satisfactory choice. (Wetherill et al., 1986)

We compare the resulting model in (5.5) to the model obtained by Wetherill et al. in their response surface modeling. Given the nature of the data and knowledge concerning the physical properties of the experiment, the investigators expected a quadratic model with interaction terms. Therefore, based on process knowledge and examination of the relationships between the predictor variables and the predictor variables with the response variable, Wetherill et al. chose to fit a quadratic model with 35 terms, including the seven squared terms and 21 cross-product variables. In general, the model he chose to fit to the data was of the form:

$$\hat{y} = b_o + \sum_{i=1}^{p} b_i x_i + \sum_{i=1}^{p} \sum_{j=i}^{p} b_{ij} x_i x_j + \sum_{k=p+1}^{2p} \sum_{i=1}^{p} b_k x_i^2 \qquad (5.6)$$

Fitting this model to the data, we obtain an ICOMP value of 886.33 for the model in (5.6). One reason for the substantial increase in the

ICOMP value for the fitted model (5.6) was the fact that only two-variable, first-order interaction terms were included in the model. From (5.5), we see that three of the terms in the model were two-variable quadratic interactions.

5.4 Discussion and Conclusion

As the computational limitations of technology has been reduced, more attention has been given recently to automatic model generation, also referred to as self-organizing modeling. Mueller (1998) recently covers various methods incompassed in the area of automatic model generation. Much of the current literature focuses on methods for selecting models from data and the concepts behind the algorithms used for the self-organization of data. (Muller et al., 1997, Ivakhnenko et al., 1994, and Fayyad, 1996) Few recent works have attempted to directly discuss the potential weaknesses of the GMDH and to develop methods for accommodating these weaknesses.

Our application of the GA, with information-based fitness functions, on the final backtraced model from the GMDH algorithm

has performed well with respect to the identification of a "better", more parsimonious submodel. For large data sets with large number of predictor variables, the GMDH algorithm tends to move to higher levels, generating extremely complex models with hundereds and thousands of terms. As seen in example using the Thackray data (Section 5.3), the GA allowed us to identify a better-fitting, less complex, and more parsimonious model to the data. The value of the GMDH lied in the ability to generate a model without assuming the functional form prior to modeling. As seen in comparing (5.4) with (5.5), the functional form of the model assumed by Wetherill et al. (5.6) had a much greater number of terms and still did not include some key, significant terms. We again note the value of being able to relax the assumption of the functional form prior to the modeling of the data. Through the use of the GMDH, a self-organizing modeling heuristic, and the application of the GA, a much more parsimonious, less complex and "better fitting" model was obtained for the data.

CHAPTER 6

THE PERFORMANCE OF THE GMDH WITH INFORMATION-BASED MODEL EVALUATION CRITERIA: EMPIRICAL RESULTS

The methods developed in this thesis are designed to contribute to the areas of statistical modeling and response surface methodologies by making self-organization methods more reliable, efficient, and effective. In order to understand the overall contributions of our work, certain specific areas were investigated. First, we claim that our method outperforms that of Ivakhnenko's algorithm with respect to two primary issues -- the elimination of the need for the subdivision of the data and better performance of information-based criteria in comparison to the use of the regularity criterion (2.4). Therefore, we compare the performance of the Ivakhnenko algorithm to the GMDH with information-based criteria using data generated from a known relationship (i.e., the functional form is known prior to the modeling of the data). We also show a

comparison of the overall performance of the GMDH under AIC (4.7), ICOMP (4.8) and ICOMPIFIM (4.20).

6.1 The Simulation Protocol

In order to assess the performance of the proposed approach presented in this thesis, we adopt a specific simulation protocol to be used across all scenarios. This protocol is used to compare the performance of AIC, ICOMP, and ICOMPIFIM as derived in Chapter 4 for the GMDH algorithm. We then compare the performance of the GMDH with Information-Based Model Evaluation Criteria to the performance of the Ivakhnenko's algorithm using this same protocol. Finally, we assess the performance of the genetic algorithm on the backtracked model as well as for pre-screening of variables prior to invoking the GMDH algorithm.

The data for each of the simulation studies are generated as follows. The set of five (5) predictor variables introduced into the algorithms are generated as:

$$x_1 = 10 + \varepsilon_1;$$
$$x_2 = 10 + 0.3\varepsilon_1 + \alpha\varepsilon_2;$$
$$x_3 = 10 + 0.3\varepsilon_1 + 0.5604\,\alpha\varepsilon_2 + 0.8282\,\alpha\varepsilon_3;$$
$$x_4 = -8 + 0.5\,\varepsilon_4;$$
$$x_5 = 5 + 0.3\,\varepsilon_5,$$

where $\alpha = \sqrt{1-(0.3)^2} = \sqrt{0.91}$ and $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$, and ε_5, are normally distributed random variables with mean zero (0) and variance one (1). The response variable, y, is generated from the following functional relationship:

$$y = -8 + x_1 + 0.5x_2 + 0.3x_3 + 0.5\varepsilon, \tag{6.1}$$

where $\varepsilon \sim N(0,\sigma)$. Obviously, for each of the simulations, it is expected that the algorithms choose a model containing x_1, x_2, and x_3.

6.2 A Comparison of the Performance of AIC, ICOMP, and ICOMPIFIM within the GMDH

In Chapters 3 and 4, three information-based model evaluation, criteria, AIC, ICOMP and ICOMPIFIM, were discussed and derived for

the GMDH environment under the assumption of normally distributed error terms. In order to understand the relative performance of the GMDH algorithm with information-based criteria used to evaluate and screen models at each level, we use the simulation protocol of Section 6.1. The response variables are generated from (6.1) with a standard deviation of σ=0.05. Table 6.1 presents summary results for 50 runs of the GMDH for each of the three information-based criteria.

We expect that the algorithm will choose a Level 2 model in x_1, x_2, and x_3 due to the nature of the GMDH algorithm. We recall the general form (2.2) of the basic polynomial used at each generation of the GMDH:

$$y=A + Bx + Cy + Dx^2 + Ey^2 + Fxy.$$

At level one, the algorithm only constructs polynomials in two variables. Thus, in order to obtain a model in three variables, the algorithm must move to level two. Since the response variable generated from (6.1) is a first-order model in x_1, x_2, and x_3, we expect the algorithm to stop at level 2.

TABLE 6.1. Results from 50 Simulation Runs of the GMDH with AIC, ICOMP, and ICOMPIFIM.

Level of Chosen Model	Variables in Chosen Model	AIC	No. Times Chosen by ICOMP	ICOMPIFIM
2	x_1,x_2,x_3	4	18	30
3	x_1,x_2,x_3	0	0	2
3	x_1,x_2,x_3,x_4	8	18	8
3	x_1,x_2,x_3,x_5	6	14	9
4	x_1,x_2,x_3,x_4,x_5	32	0	1

The results summarized in Table 6.1 contain the frequency of times that the GMDH algorithm chooses the "best approximating model" to be a model of the level in column 1 containing the variables in column 2. These initial simulation runs clearly demonstrate that the GMDH algorithm performs poorly in the case of Akaike's Information Criterion (AIC). In Chapter 4, it was suggested from the numerical example on the Wakely (1949) data that AIC tended to "overfit" the model. We recall the form of AIC as derived for the GMDH algorithm in (4.7):

$$AIC(GMDH_{normal}) = n\log(2\pi) + n\log(\hat{\sigma}) + n + 2k,$$

where n represents the number of observations in the data set and k is the number of parameters estimated in the model. Due to the nature of the GMDH algorithm, AIC fails to punish for the growing complexity of the model as the algorithm moves from level-to-level. Therefore, in the GMDH environment, AIC chooses to add terms to the model in order to obtain a "better fitting" model (a weakness common to the coefficient of determination, R^2, in regression analysis). This weakness is clearly seen in the results of Table 6.1, as AIC chooses a higher-order model in all five predictor variables 64% of the time. AIC does not appropriately distinguish model complexity and lack of parsimony in the GMDH environment.

The results in Table 6.1 also suggest that the information-based criterion ICOMPIFIM performs better than ICOMP in the GMDH algorithm. The inherent structure of the inverse-Fisher information matrix punishes for lack of parsimony as well as for added complexity through the incorporation of both the error variance and

the $(X'X)^{-1}$ matrix in the evaluation criterion. This result will be examined in further simulation studies, comparing the performance of ICOMP and ICOMPIFIM in the GMDH with the performance of the regularity criterion (2.4) in the algorithm.

6.3 A Comparison of the Performance of the GMDH with Information-Based Criteria to the Original Ivakhnenko Algorithm

Three primary weaknesses of Ivakhnenko's algorithm were discussed in Chapter 2. The first weakness relates to the subdivision of the data set into the training and checking sets. The subdivision of the initial data set reduces the amount of information available for model building. Also, the approach to subdividing the data randomly can yield substantially different results for the same set of data. The second major weakness of the original GMDH was the overall performance of the regularity criterion (2.4). Basically, the regularity criterion is the root mean squared error computed as the difference between the predicted values and the actual values for the values of the observations in the checking set. Thus, the

regularity criterion has a tendency, as do other criteria based solely on mean squared error, to choose overparameterized models. The fact that the value of the regularity criterion is computed from the observations in the checking set renders it a weak criterion for "noisy" data.

The incorporation of information-based model evaluation criteria in the GMDH was designed to eliminate these three primary weaknesses from the algorithm. The objective is to obtain a self-organizing modeling algorithm that

- does NOT require subdivision of the data;
- does NOT tend to choose overparameterized models; and that
- DOES perform well for "noisy" data sets (i.e., for data sets from processes that exhibit higher amounts of variation).

Empirical studies show that the proposed algorithm performs well in overcoming each of these stated weaknesses.

In order to illustrate the ability of our algorithm to overcome the weaknesses of the original GMDH, Monte Carlo simulation studies were performed to compare our algorithm under two different levels of variation in the response variable, y. Tables 6.2 and 6.3 provide

TABLE 6.2 Monte Carlo Experimental Results for GMDH with σ=0.05 (100 Experimental Runs)

Level	Variables in Model	Frequency Original	Frequency ICOMP	Frequency ICOMPIFIM
1	x_1,x_2	5	0	0
2	x_1,x_2,x_3	10	38	64
2	x_1,x_2,x_3,x_4	1	0	0
3	x_1,x_2,x_3	8	0	3
3	x_1,x_2,x_3,x_4	26	35	16
3	x_1,x_2,x_3,x_5	19	27	16
3	x_1,x_2,x_3,x_4,x_5	31	0	0
4	x_1,x_2,x_3,x_4	0	0	1

TABLE 6.3 Monte Carlo Experimental Results for GMDH with σ=0.25 (100 Experimental Runs)

Level	Variables in Model	Frequency Original	Frequency ICOMP	Frequency ICOMPIFIM
1	x_1,x_2	22	0	0
1	x_1,x_3	4	0	0
2	x_1,x_2,x_3	11	39	61
2	x_1,x_2,x_5	1	0	
3	x_1,x_2,x_3	4	0	6
3	x_1,x_2,x_3,x_4	8	30	10
3	x_1,x_2,x_3,x_5	12	29	17
3	x_1,x_2,x_4,x_5	2	0	0
3	x_1,x_2,x_3,x_4,x_5	7	0	2
4	x_1,x_2,x_3,x_4	5	0	1
4	x_1,x_2,x_3,x_5	4	1	0
4	x_1,x_2,x_3,x_4,x_5	20	1	3

summaries of the results for 100 simulation runs for each of the three algorithms:

1) The GMDH with ICOMP criterion;

2) The GMDH with ICOMPIFIM criterion; and

3) The original GMDH with the regularity criterion.

6.3.1 Empirical Results with Respect to Model Identification

For the results shown in Table 6.2, the observations were generated according to the protocol in Section 6.1 with σ=0.05. Table 6.3 contains the summarized results for the situation in which σ=0.25 (i.e., the variability in the responses is much larger). For both tables, the first column represents the level that contains the model with the minimum value of the evaluation criterion (e.g., regularity, ICOMP or ICOMPIFIM). The second column contains the variables that are contained in the backtraced model chosen by the algorithm. The last columns represent the number of times under ICOMP, ICOMPIFIM and the original algorithm, respectively, that the

algorithm chooses a model from the specified level with the specified variables.

From the results in Table 6.2, it is obvious that our proposed algorithm performs much better than the original algorithm, even in the situation of low variability. The GMDH with the ICOMP criterion chooses the correct model (i.e., a level 2 model in x_1,x_2, and x_3) in 38% of the experimental runs; whereas, the original algorithm only chose the best model 10% of the time. The proposed algorithm performs even better using the information-based criterion ICOMPIFIM. The GMDH with ICOMPIFIM chose the correct model 64 out of 100 runs -- a 600% improvement over the original algorithm!

For the situation in which the variation is increased to σ=0.25, the GMDH with information-based model evaluation criteria performs almost identically to its performance under σ=0.05. We see in Table 6.3, that the GMDH with ICOMP chooses the correct model 39% of the time in the situation of higher variability -- almost identical to its performance under lower variability. Similarly, the GMDH with ICOMPIFIM chooses the correct model 61% of the time when σ=0.25 vs. 64% when σ=0.05. On the other hand, the original

algorithm had greater difficulty identifying the appropriate level in which to stop under the situation with greater variability. The original algorithm still chose the best model on 11 out of 100 runs. However, it also chose a first level model in only two variables 22% of the time vs. 5% of the runs in the case with σ=0.05. More critically, the original algorithm never chose a fourth level model in the case of lower variation. When the "noise" in the data was increased by 5x, the original GMDH chose a fourth level model 29% of the time. Hence, it had a much more difficult time approaching the correct model in the situation of increased variation. The GMDH with information-based criteria was fairly robust to the increased variability, with the results remaining similar across the two levels of variation.

6.3.2 Empirical Results with Respect to Algorithm Efficiency

One of the challenges faced by the entire field of self-organizing modeling methods and automatic model generation techniques is the computational burdens that result due to the large

number of terms and models that are evaluated. Thus, a major consideration of the performance of the proposed algorithm is the efficiency in which the algorithm progresses until it reaches the stopping criterion. We define efficiency in this framework in terms of the total number of models that the algorithm must evaluate prior to terminating, either by attaining the "best" approximating model or by reaching the maximum number of allowed generations. The efficiency of the GMDH with information-based criteria is compared to the efficiency of the Ivakhnenko algorithm by identifying the number of models evaluated by the algorithm in 50 simulation runs of each algorithm for the protocol defined in Section 6.1 under the cases of σ=0.05 and σ=0.25.

Figure 6.1 shows the frequency distribution of the number of models evaluated under the original algorithm for the two levels of variability. Obviously, as the "noise" in the data increases, the range of the number of models evaluated increases drastically. For the case of σ=0.05, the maximum number of models evaluated by the algorithm in one experimental run was 389. However, when σ was increased to 0.25, the algorithm was required to evaluate up to 5,281

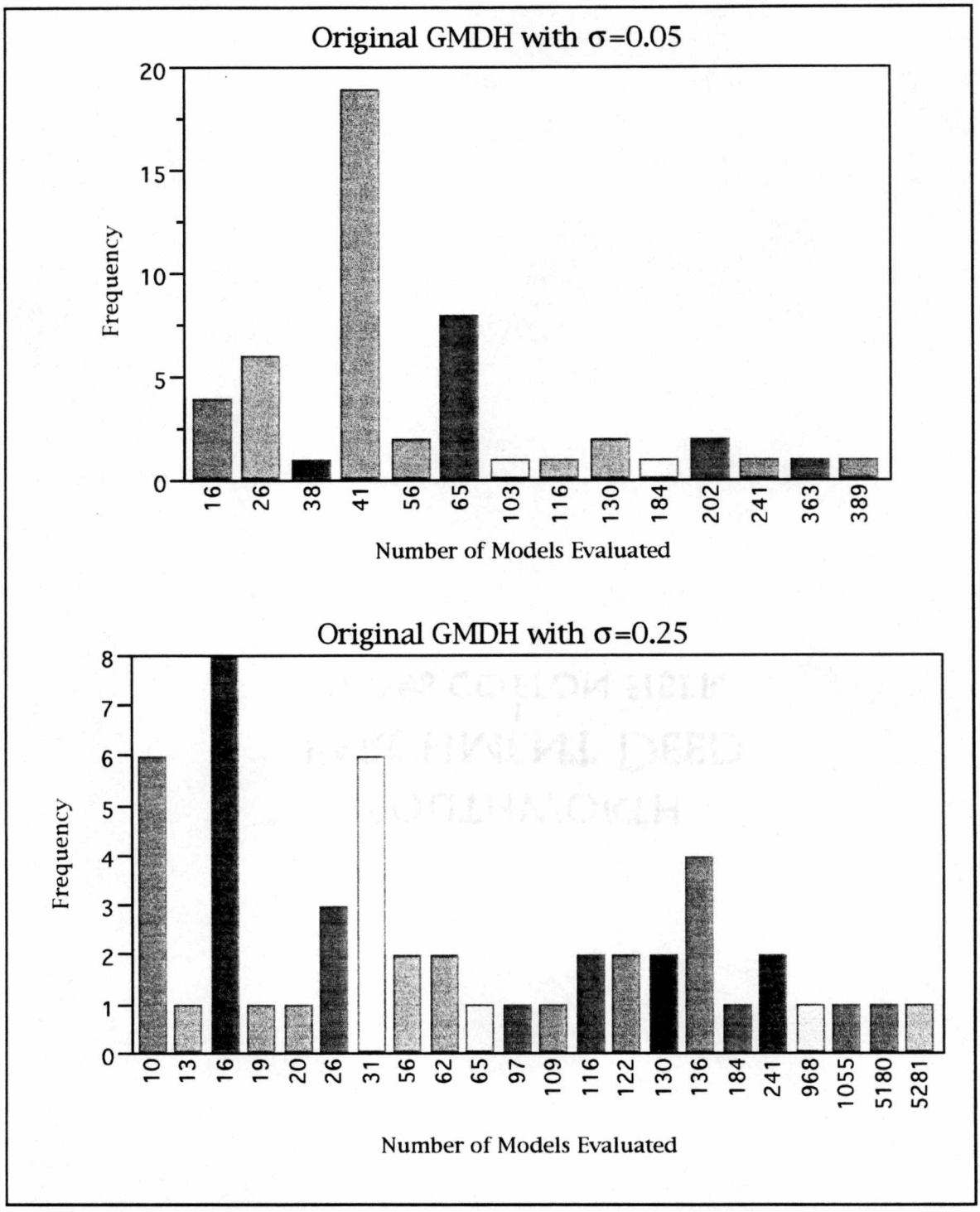

FIGURE 6.1. Frequency Distributions for the Number of Models Evaluated by the Original GMDH for σ=0.05 and σ=0.25

models -- a much greater computational burden. Figures 6.2 and 6.3 show the efficiency of the GMDH with ICOMP and ICOMPIFIM, respectively. As seen in these figures, the GMDH with information-based criteria consistently requires the evaluation of fewer models to identify the correct model. Even with the increased variability in the response variable, the proposed algorithm performs much more efficiently. Under ICOMP, the maximum number of models the algorithm was required to evaluate was 146. With ICOMPIFIM, the maximum number evaluated was 96. Thus, the proposed algorithm consistently performs much more efficiently than Ivakhnenko's algorithm.

6.4 Performance of the GA in the GMDH Environment

Given the structure of the polynomials used to construct the polynomials at each level of the GMDH, the model chosen by the algorithm contains terms that are insignificant in terms of the predictive power of the model. We therefore utilize the GA as a means to "backfit" the final GMDH model to a more parsimonious

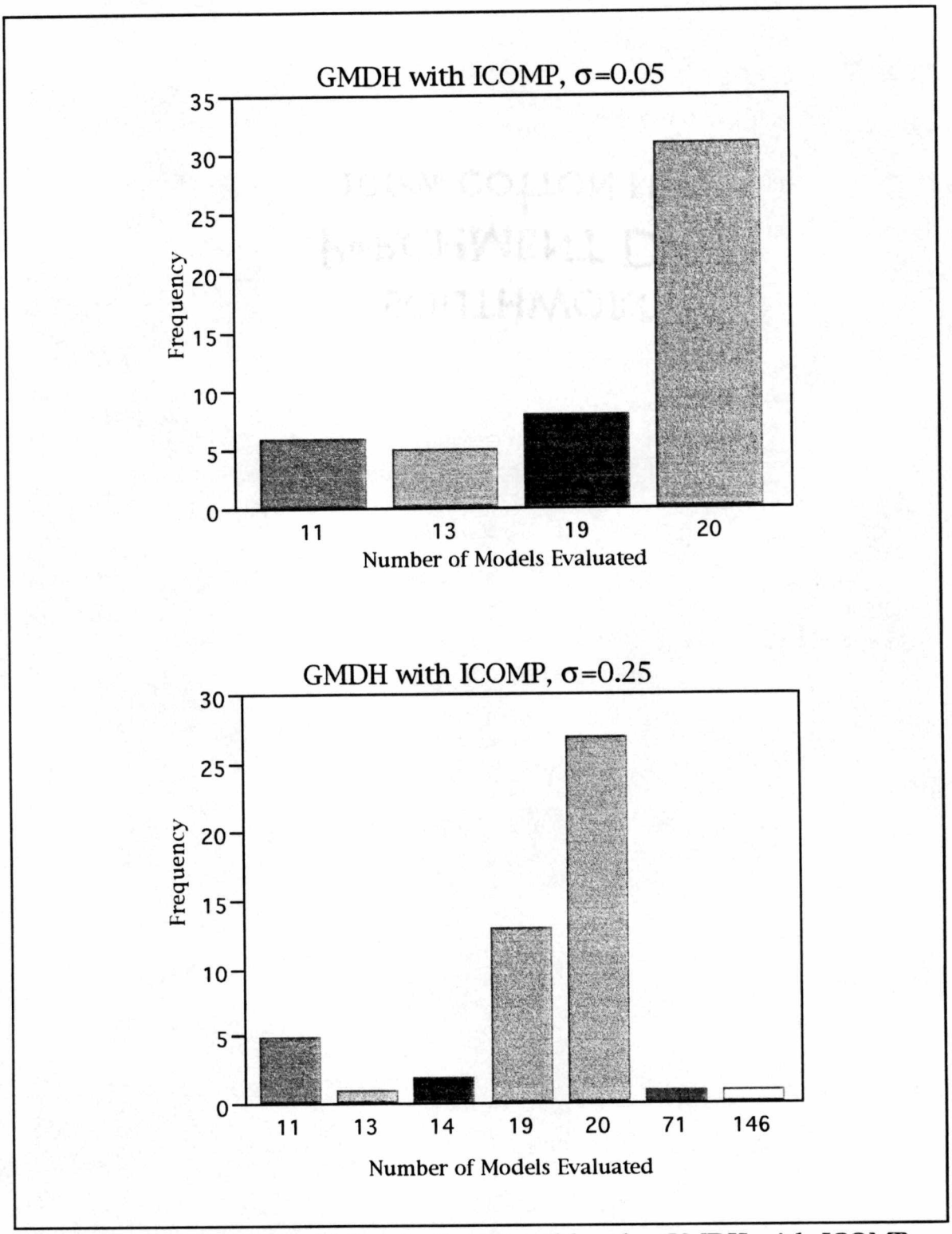

FIGURE 6.2. Number of Models Evaluated by the GMDH with ICOMP for σ=0.05 and σ=0.25

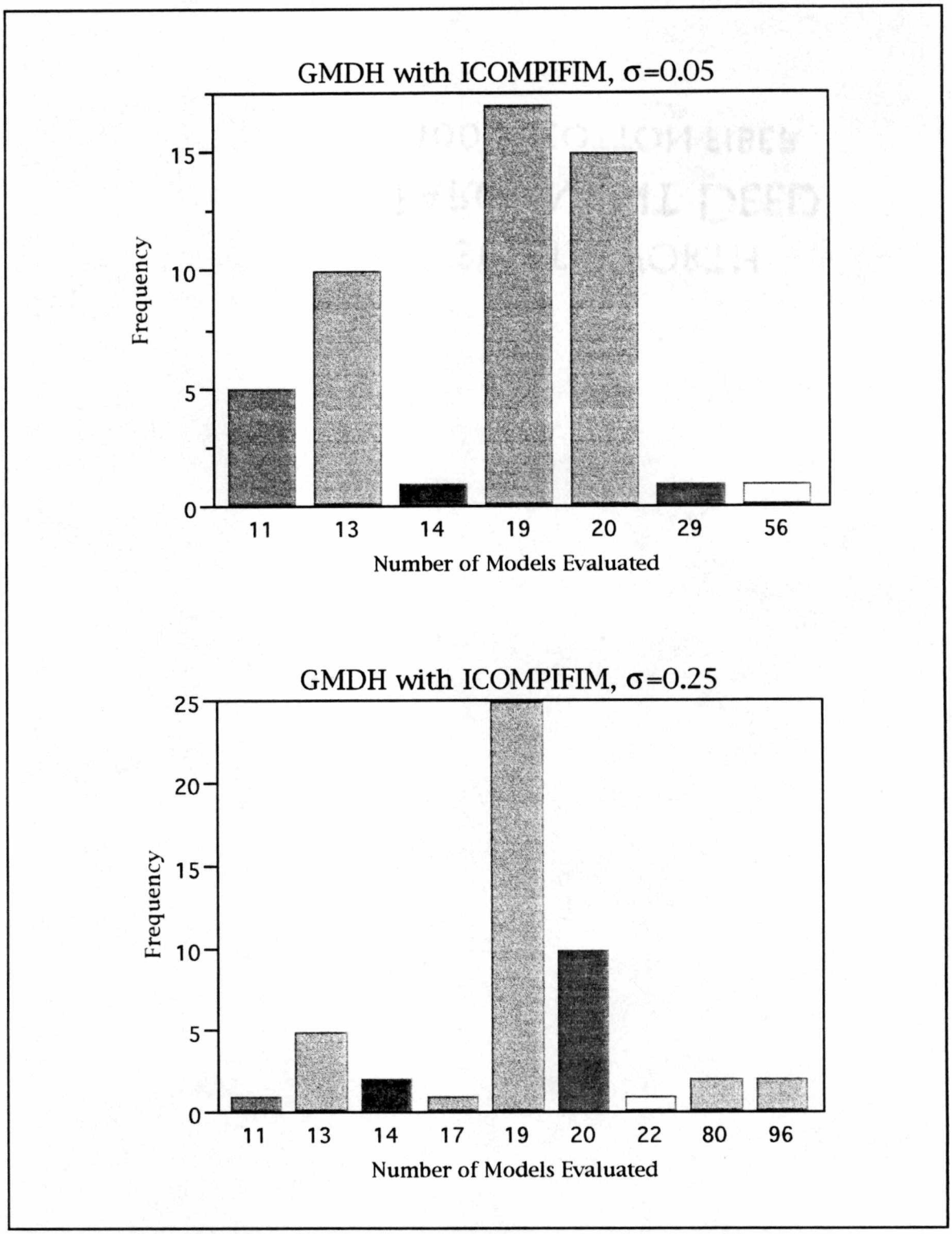

FIGURE 6.3. Number of Models Evaluated by the GMDH with ICOMPIFIM for σ=0.05 and σ=0.25

form. There still exist several performance questions with the application of the GA dealing with the required number of generations, the selection of the population size, and the choice of mutation and crossover probabilities.

In order to understand the ability of the GA to effectively "backfit" the final GMDH model, we ran the GA under various population sizes and number of generations with probability of crossover = 0.75 and probability of mutation = 0.01. These experimental runs were performed based on the level 2 GMDH model in x_1, x_2, and x_3. Thus, the input into the GA is the set of terms in the backtraced GMDH model shown in Table 6.4.

Table 6.5 provides a summary of the performance of the GA for under various population sizes, n, and number of generations. As illustrated, the GA performs better with the larger population sizes (n=75) across the various number of generations that the algorithm was run. The resulting values for ICOMPIFIM(GA) reported in Table 6.4 can be compared to the value for the "best" model from which the data were generated. In this case, the model with terms {constant, x_1, x_2, and x_3} has an ICOMPIFIM criterion value of -353.9.

TABLE 6.4 Terms in the Backtraced GMDH Model

Term No.	Variables in Term	Term No.	Variables in Term
0	x_1	17	$x_1x_3^2$
1	x_2	18	$x_2x_3^2$
2	x_3	19	$x_1x_2^3$
3	x_4	20	$x_1x_3^3$
4	x_1^2	21	$x_1^2x_2$
5	x_2^2	22	$x_1^2x_3$
6	x_3^2	23	$x_1^2x_2^2$
7	x_1^3	24	$x_1^2x_3^2$
8	x_2^3	25	$x_1^3x_2$
9	x_3^3	26	$x_1^3x_3$
10	x_1^4	27	$x_2^2x_3$
11	x_2^4	28	$x_2^2x_3^2$
12	x_3^4	29	$x_1x_2x_3$
13	x_1x_2	30	$x_1x_2x_3^2$
14	x_1x_3	31	$x_1x_2^2x_3$
15	x_2x_3	32	$x_1^2x_2x_3$
16	$x_1x_2^2$		

TABLE 6.5 Results on the Backfitting of the GMDH Model via the Genetic Algorithm

n	# Generations	ICOMPIFIM	Terms in Model
50	50	-310	1,2,4,6,13,14,15, 17,21,25
50	75	-323.9	0,1,2,6,14,17
50	100	-338.4	0,2,3,4,7
50	200	-320	0,1,2,6,8,9
75	50	-338	0,1,2,3,8,12
75	75	-333.1	0,1,2,3,6,14,17
75	100	-336	0,1,2,3,5,14

In Chapter 5, we also demonstrated the use of the genetic algorithm for pre-screening variables prior to invoking the GMDH algorithm. The ability of the GA with information-based fitness functions to identify the "best" subset of predictors was studied by performing 100 runs using the protocol described in Section 6.1. The input into the GA was the variable set {constant,x_1,x_2,x_3,x_4,x_5}. Recalling the function in (6.1) from which the values for the response variables were obtained, we expect that the GA consistently identifies either the subset {constant,x_1,x_2,x_3} as the set of variables

that should be entered into the GMDH algorithm. The GA is run under the following conditions: n=6, no. of generations = 25, p(crossover)=.75 and p(mutation)=.01 with ICOMPIFIM as the fitness function. The results showed that the GA chose the correct subset of variables to enter into the GMDH in 81 out of the 100 runs. (Obviously, this proportion can be improved by increasing the number of generations and/or the population size, n.) Thus, 81% of the time, the GA selected only the variables contained in the correct model. The other 11% of the time the GA chose either {constant,x_1,x_2,x_3,x_4} or {constant,x_1,x_2,x_3,x_5}.

As discussed in Chapter 7, the use of the GA is a means to compensate for one of the primary weaknesses of the GMDH algorithm--its tendency to build unnecessarily complex models. However, through backfitting with the GA, we can obtain a much more parsimonious model that maintains the goodness-of-fit, improves the parsimony, and decreases the complexity. Again, the value in the overall approach is its ability to "self-organize" a functional model of a set of data with a relatively small number of observations and without assuming any functional form prior to

modeling. Through pre-screening of variables using the GA, we are able to reduce the computational burden of the GMDH algorithm.

CHAPTER 7

FUTURE RESEARCH ON THE GMDH AND RELATED AREAS

The GMDH is a statistical modeling heuristical approach that attempts to deal with the common problem of lack of knowledge concerning the proper structure of a model. For example, in stock identification modeling, it is typically assumed that the best functional form for the modeling of the data is either linear or quadratic. This assumption is often questionable or known to be false. (Prager, 1988) However, most modeling and response surface methodologies have focused on the estimation of the parameters of a given model and not to determine the functional form of the model.

7.1 Strengths and Limitations of the GMDH

The idea behind the self-organizing modeling methods is that the functional form of the model can be determined from the data, not from assumptions of the researcher. The GMDH algorithm is a heuristical method that attempts to identify a higher-order polynomial that optimizes a model evaluation criterion (in our case, either ICOMP or ICOMPIFIM). With this heuristic, it is desirable that the investigator only supplies the data set into the modeling technique. Any other information is obtained from the observations themselves. The GMDH as developed in this thesis has several advantages:

- It allows for the fitting of a high-order model without a large amount of data. Given the iterative, hierarchical nature of the algorithm, a complete multinomial model can be estimated with less than 100 data observations.
- The algorithm eliminates the need for subdivision of the data into a "training" and "checking" set, allowing for the full information

available in the observations to be used in the model building stages of the algorithm.

- The algorithm using information-based criteria performs extremely well in the face of "noisy" data sets. The information-theoretic model evaluation criteria, ICOMP and ICOMPIFIM, are fairly robust to reasonable increases in the variation exhibited by the data. Poor performance in the face of "noisy" data has been one of the major criticisms of self-organizing modeling methods.
- The developed algorithm is computationally more efficient than previous forms of self-organization, data handling techniques, requiring the evaluation of fewer models in order to minimize the error statistic, ICOMP or ICOMPIFIM.
- The algorithm does provide a "good" estimated model in situations where the researcher/modeler has limited knowledge on the behavior of a system prior to data collection. There is no claim by ourselves, by Ivakhnenko, or by other researchers in this area that self-organizing methods provide the optimal structure -- "merely a good one". (Ivakhnenko et al., 1979, Farlow, 1981, and Prager, 1988)

As with many heuristical methods, the approach as presented in this thesis still has several weaknesses. The primary weakness is that in order to identify a "good" model, the algorithm often builds a full polynomial of higher-order that contains many terms that need not be used. The second major weakness lies in the fact that the method still assumes that the error terms are normally distributed. Finally, the size of the models being built at higher levels of the algorithm still create a huge computational burden. The nature of the areas which still need to be studied in the area of self-organizing modeling methods are based on these three primary weaknesses. Therefore, our future research in the area of the GMDH is focused on relaxing the assumption of normality, reducing the unnecessary complexity from the models generated by the GMDH, and working to reduce the computational burden that the methodology poses.

7.2 Future Research Endeavors

The current algorithm is concerned with the identification of a functional form to model a set of data with minimal assumptions

required by the researcher. One major assumption still remains -- normally distributed error terms. Future work is to be concerned with the selection of the functional model and the estimation of the parameters in the situation of generalized linear models (GLM's), a class of models introduced by Nelder and Wedderburn (1972). Since the class of GLM's is large, encompassing ordinary linear regression, logistic regression, Poisson regression, etc., it is undesirable to derive the form of the information-based criteria for every possible distribution covered by the class of GLM's. Nordberg (1981) shows that he is able to perform variable selection for GLM's by converting the problem into a variable selection problem for an ordinary linear regression model. In 1982, Nordberg presents his empirical studies which shows that a quadratic log likelihood approximation works well for variable selection purposes. We plan to investigate the use of Nordberg's approximation of the log likelihood and his proposed data transformation for the area of model evaluation and screening within the framework of the GMDH. If his procedure can be appropriately modified, it can serve as a means to relax the

assumption of normality without altering the computational structure of the GMDH.

Even though our approach has effectively reduced the number of models being evaluated by the algorithm, the GMDH still produces models with terms that are not useful or significant with respect to improving the predictive power of the algorithm. We have attempted to address this weakness by using the GA to "backfit" the final model obtained from the GMDH. However, the GA is also computationally expensive, evaluating a large number of models at each generation within the GMDH environment. Since a level four model in six variables could contain up to 7,200 terms, the population size for the GA would need to be quite large (e.g., n=500 or n=1,000) in order to cover such a large potential response surface. We plan to investigate structural changes within the GMDH to eliminate the need for extensive "backfitting" after model identification and estimation. Tamura and Kondo (1980) have proposed the identification of optimal partial polynomials at each level and only passing the partial polynomials to the next stage. As a branch of Tamura and Kondo's work, we plan to investigate

modifications to the quadratic polynomials in the current algorithm while maintaining ICOMP and ICOMPIFIM as the model evaluation functions.

Empirical evidence shows that the developed approach handles the problem of model identification and estimation for problems with noise five times greater than those handled by the original algorithm. We also desire to investigate the predictive power of the GMDH with information-theoretic criteria for modeling problems with noise 10 to 25 times greater than the levels currently being investigated in the literature. According to Ivakhnenko (1984), this task is "necessary in order to solve problems in the prediction of complex processes in ecology, meteorology, economics, and other areas". Some attempts have been made to accomplish this task by incorporating pattern recognition and coding theories into the GMDH. (Yurachkovskiy, 1982, Ivakhnenko et al., 1981) These attempts have primarily focused on deterministic prediction (Ivakhnenko et al, 1982), not on the development of evaluation criteria which are increase the GMDH's robustness to noise in the systems being modeled.

REFERENCES

REFERENCES

----------, *Tax Research: IRS has made Progress but Major Challenges Remain,* Report to the Commissioner, Internal Revenue Service/U.S. General Accounting Office, Washington, D.C.: The Office, Gaithersburg, MD, 1996.

Akaike, H., *Information Theory and an Extension of the Maximum Likelihood Principle,* Proceedings of the Second International Symposium on Information Theory, B. Petrov and F. Csaki (eds.), 1973, Budapest: Academiai Kiado, 267-281.

Akaike, H., *A New Look at the Statistical Model Identification,* IEEE Transactions on Automation and Control, 1974, AC-19, 716-723.

Akaike, H., *On Entropy Maximization Principle,* Proceedings of the Symposium on Application of Statistics, P. R. Krishnaiah (ed.), Amsterdam, North Holland, 1977.

Akaike, H., *Likelihood of a Model and Information Criteria,* Journal of Econometrics, 1981, 16, 3-14.

Anderson, R. L. and Bancroft, T. A., Statistical Theory in Research, NY: McGraw-Hill, 1959.

Barron, R. L., Mucciardi, A. N., Cook, F. J., Craig, J. N., and Barron, A. R., *Adaptive Learning Networks: Development and Applications in the United States of Algorithms related to GMDH,* Self-Organizing Methods in Modeling, S. J. Farlow (ed.), NY: Marcel Dekker, 1984, 25-65.

Bearse, P. M. and Bozdogan, H., *Selecting the Model Order, Lag Length, and the Best Predictors in Vector Autoregressive (VAR) Models,* SAMS, 1998, 31(1-2), 61-91.

Behboodian, J. *Information for Estimating the Parameters in Mixtures of Exponential and Normal Distributions,* Ph.D. Thesis, Dept. of Mathematics, University of Michigan, Ann Arbor, 1964..

Boltzmann, L., *Uber die Beziehung zwischen dem zweitin Hauptsatze der Mechnaischen Warmetheorie und der Wahrscheinlichkeitsrechnung respective den Satzen uber das Warmegleichgewicht,* Wiener Berichte, 1877, 76, 373-435.

Bowerman, B. L., O'Connell, R. T., and D. A. Dickey, Linear Statistical Models: An Applied Approach, Boston: Duxbury Press, 1986.

Boyce, D. E., Farhi, A. and Weischedel, R., Optimal Subset Selection: Multiple Regression, Interdependence, and Optimal Network Algorithms, NY: Stringer-Verlag, 1974.

Bozdogan, H., *Multi-Sample Cluster Analysis and Approaches to Validity Studies in Clustering Individuals,* Ph.D. Thesis, The University of Illinois at Chicago Circle, 1981.

Bozdogan, H., *Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extension,* Psychometrika, 1987, 52(3), 345-370.

Bozdogan, H., *ICOMP: A New Model-Selection Criterion,* Classification and Related Methods of Data Analysis, H. H. Bock (ed.), Amsterdam: Elsevier Science Publishers (North Holland), 1988, 599-608.

Bozdogan, H., *On the Information-Based Measure of Covariance Complexity and Its Application to the Evaluation of Multivariate Linear Models,* Communications in Statistics Theory and Methods, 1990, 19(1), 221-278.

Bozdogan, H., *Mixture-Model Cluster Analysis Using a New Informational Complexity and Model Selection Criteria*, Multivariate Statistical Modeling, Vol. 2, H. Bozdogan (ed.), Dordrecht, Netherlands: Kluwer Academic Publishers, 1994.

Bozdogan, H., and Haughton, P. M. A., *Informational Complexity Criteria for Regression Models*, Computational Statistics and Data Analysis, 1996.

Bradbury, R. H., Hammond, L. S., Reichelt, R. E., and Young, P. C., *Prediction versus Explanation in Environmental Impact Assessment*, Search, 1984, Vol. 14, No. 11-1, 323-325.

Chernoff, H., *Large Sample Theory: Parametric Case*, Annals of Mathematical Statistics, 1956, 27, 1-22.

Claereboudt, M. R., *GMDH Algorithm as a Tool for Bivalve Growth Analysis and Prediction*, ICES Journal of Marine Science, 1994, Vol. 51, No. 4, 439.

Cleran, Y., Thibault, J., Cheruy, A., and Corrieu, G., *Comparison of Prediction Performances Between Models Obtained by the Group Method of Data Handling and Neural Networks for the Alcoholic Fermentation Rate in Enology*, Journal of Fermentation and Bioengineering, 1991, Vol. 71, No. 5, 346-362.

Draper, N. R. and Smith, H., Applied Regression Analysis, NY: John Wiley & Sons, 1981.

Farlow, S. J., *The GMDH Algorithm of Ivakhnenko*, American Statistician, 1981, Vol., 35, No. 4, 210-215.

Farlow, S. J., *The GMDH Algorithm*, Self-Organizing Methods in Modeling, S. J. Farlow (ed.), 1984, 1-24.

Fayyad, U. M., *From Data Mining to Knowledge Discovery: An Overview*, Advances in Knowledge Discovery and Data Mining, U. M. Fayyad (ed.), MIT Press, 1996.

Fisher, R. A., *On an Absolute Criterion for Fitting Frequency Curves*, Mess. of Math., 1912, 41, 155.

Fisher, R. A., *On Mathematical Foundations of Theoretical Statistics*, Philosophical Transactions of the Royal Society A, 1921, 222, 309.

Fisher, R. A., *Theory of Statistical Estimation*, Proceedings of the Cambridge Philosophical Society, 1925, 22, 700.

Fisher, R. A., *Two New Properties of Mathematical Likelihood*, Proceedings of the Royal Society, London, 1934, A-144, 285.

Fisher, R. A., *The Logic of Inductive Inference*, Journal of the Royal Statistical Society, London, 1935, 98, 39-54.

Fisher, R. A., Statistical Methods and Scientific Inference, London: Oliver and Boyd, 1956.

Goldberg, D. E., Genetic Algorithms in Search, Optimization, and Machine Learning, NY: Addison-Wesley, 1989.

Green, D. G., Bradbury, R. H., Reichelt, R. E., *Patterns of Predictability in Coral-Reef Community Structure*, Coral Reefs, 1987, Vol. 6, No. 1, 27-34.

Green, D. G., Reichelt, R. E., Buck, R. G., *Self-Adaptive Modeling Algorithms*, Mathematics and Computers in Simulation, 1988, Vol. 30, No. 1-2, 33-38.

Green, D. G., Reichelt, R. E., Bradbury, R. H., *Statistical Behavior of the GMDH Algorithm*, Biometrics, 1988, Vol. 44, No. 1, 49-69.

Haughton, P. M. A., *Consistency of a Class of Information Criteria for Model Selection in Non-linear Regression*, Communications in Statistics-Theory and Methods, 1991, 20, 1619-1629.

Hirata, H., *Modeling and Analysis of Ecological Systems - The Large-Scale System Viewpoint*, Internation Journal of Systems Science, 1987, Vol. 18, No. 10, 1839- 1855.

Hirata, H., and Ulanovicz, R. E., *Large-Scale System Perspectives on Ecological Modeling and Analysis*, Ecological Modelling, 1986, Vol. 31, No. 1-4, 79-104.

Hocking, R. R., *The Analysis and Selection of Variables in Linear Regression*, Biometrics, 1976, 32, 1044.

Holland, J., *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, 1975.

Holland, J., *Genetic Algorithms*, Scientific American, 1992, 66-72.

Ikeda, S., Ochiai, M., and Sawaragi, Y., *Sequential GMDH Algorithm and Its Application to River Flow Prediction*, IEEE Transactions on Systems, Man, and Cybernetics, 1976, SMC-6(7), 473-479.

Ivakhnenko, A. G., *Group Method of Data Handling-A Rival of the Method of Stochastic Approximation*, Soviet Automatic Control, 1966, 13, 43-71.

Ivakhnenko, A. G. and Ivakhnenko, G. A., *Long-term Prediction by GMDH Algorithms Using the Unbiased Criterion and the Balance-of-Variables Criterion*, Soviet Automatic Control, 1974, 7(4), 40-45.

Ivakhnenko, A. G., Krotov, G. I., and Visotsky, V. N., *Identification of the Mathematical Model of a Complex System by the Self-Organization Method*, Theoretical Systems Ecology: Advances and Case Studies, E. Halfon (ed.), NY: Academic Press, 1979.

Ivakhnenko, A. G., Ivakhnenko, G. A., and Muller, J. A., *Self Organization of Neuronets with Active Neurons*, Pattern Recognition and Image Analysis, 1994, 4(2), 177-188.

Ivakhnenko, A. G., Muller, J. A., *Present State and New Problems of Further GMDH Development*, Systems Analysis, Modelling, Simulation, 1995, Vol. 20, No. 1-2, 3.

Kullback, S. and Liebler, R. A., *On Information and Suffering*, Annals of Mathematical Statistics, 1951, 22, 79-86,

Kullback, S., Information Theory and Statistics, NY: John Wiley & Sons, 1959.

Lange, T., *Structure Criteria for Automatic Model Selection in Multilayered GMDH Algorithms in Case of Uncertainty of Data*, Systems Analysis, Modelling, Simulation, 1995, Vol. 20, No. 1-2, 79.

Lehman, E. L., Testing Statistical Hypotheses, NY: John Wiley, 1959.

Lin, C. T. and Lee, C. S. G., Neural Fuzzy Systems, Upper Saddle River: Prentice Hall, 1996.

Lin, Z. S., Liu, J., and He, X. D., *The Self-Organizing Methods of Long-Term Forecasting (I) - GMDH and GMPSC Model*, Meteorology and Atmospheric Physics, 1994, Vol. 53, No. 3/4, 155.

Lindgren, B. W., Statistical Theory, Third Edition, New York: McMillan Publishing Co., Inc., 1976.

Luh, H.-K., Minesky, J. J., and Bozdogan, H., *Choosing the Best Predictors in Regression Analysis via the Genetic Algorithm with Informational Complexity as the Fitness Function*, under review in Data Mining and Knowledge, 1988.

Mantel, N., *Why Stepdown Procedures in Variables Selection*, Technometrics, 1970, 12, 591-612.

Miller, A. J., Subset Selection in Regression, London: Chapman and Hall, 1990.

Moses, L. E., Think and Explain with Statistics, Reading: Addison-Wesley, 1986.

Muller, J. A., Lemke, F., and Ivakhnenko, A. G., *GMDH Algorithms for Complex Systems Modelling*, Mathematical Modelling of Systems, 1997.

Nagasaka, K., Ichihashi, H., and Leonard, R., *Neuro-fuzzy GMDH and its Application to Modelling Grinding Characteristics*, International Journal of Production Research, 1995, Vol. 33, No. 5, 1229.

Nagata, M., Takada, K., and Sakuda, M., *Nonlinear Interpolation of Mandibular Kinesiographic Signals by Applying Sensitivity Method to a GMDH Correction Model*, IEEE Transactions on Bio-Medical Engineering, 1991, Vol. 38, No. 4, 326.

Nelder, J. A. and Wedderburn, R. W. M. *Generalized Linear Models*, JRSS A, 1972, 135, 370-384.

Neter, J., Wasserman, W., and Kutner, M., Applied Linear Regression Models, Homewood, IL: Richard D. Irwin, Inc., 1983.

Nishii, R., *Asymptotic Properties of Criteria for Selection of Variables in Multiple Regression*, Annals of Statistics, 1984, 758-765.

Nishikawa, T., and Shimizu, S., *The Characteristics of a Biased Estimator Applied to the Adaptive GMDH*, Mathematical and Computer Modelling, 1993, Vol. 17, No. 1, 37.

Nordberg, L., *A Procedure for Selection of Explanatory Variables in Generalized Linear Models*, Technical Report TRITA-MAT, Royal Institute of Technology, Stockholm, 1981, 24.

Nordberg, L., *On Variable Selection in Generalized Linear and Related Regression Models*, Commun. Statist.--Theory Meth., 1982, 11(21), 2427-2449.

Nygard, K., Ficek, R. K., and Sharda, R., *Genetic Algorithms*, OR/MS Today, August, 1992, 28-34.

Ohashi, K., *GMDH Forecasting of U.S. Interest Rates*, Self-Organizing Methods in Modeling, S. J. Farlow (ed.), 1984, 199-214.

Prager, M.H., *Group Method of Data Handling - A New Method for Stock Identification*, Transactions of the American Fisheries Society, 1988, Vol. 117, No. 3, 290-296.

Ravindra, H. V., Raghunandan, M., Srinivasa, Y. G., Drishnamurthy, R., *Tool Wear Estimation by Group Method of Data Handling in Turning*, International Journal of Production Research, 1994, Vol. 32, No. 6.

Rice, H.V., *Application of the GMDH Algorithm to Climate/Crop Forecasts*, American Society of Agricultural Engineering Report, Winter Meeting, 1979, New Orleans.

Rissanen, J., *Minmax Entropy Estimation of Models for Vector Processes*, System Identification, R. K. Mehra and D. G. Lainiotis (eds.), NY: Academic Press, 1976, 97-119.

Sarychev, A. P., *The J-Optimal Set of Regressors Determination by the Repeated Samples of Observations in the Group Method of Data Handling*, Systems Analysis, Modelling, Simulation, 1995, Vol. 20, No. 1-2, 59.

Serrano, A., and Gentil, S., *Aquatic Ecosystem Identification Using the Group Method of Data Handling*, Environmental Technology Letters, 1985, Vol. 6, No. 12, 517-525.

Sforna, M., *Searching for the Electric Load-Weather Temperature Function by Using the Group Method of Data Handling*, Electric Power Systems Research, 1995, Vol. 32, No. 1, p. 1.

Shannon, C. E., *The Bandwagon*, IRE Transactions on Information Theory, 1956, IT-2, 3.

Sheiekhova, V. Yu, *Harmonic Algorithm GMDH for Large Data Volume*, Systems Analysis, Modelling, Simulation, 1995, Vol. 20, No. 1-2, 117.

Sokal, R. R. and Rohlf, F. J., Biometry, 2nd Edition, NY: W. H. Freeman and Co., 1981.

Stewart, S. Econometrics, NY: Philip Allen, 1991.

Tamura, H. and Kondo, T., *On Revised Algorithms of GMDH with Applications*, Self-Organizing Methods in Modeling, S. J. Farlow (ed.), NY: Marcel Dekker, 1984.

Van Emden, M. H., *An Analysis of Complexity*, Amsterdam: Mathematical Centre Tracts, 35.

Wakely, J.T., *Annual Progress Report on the Soils-Weather Project*, Institute of Statistics Mimeo Series, The University of North Carolina in Raleigh, 19, 1949.

Wang, X. F., and Liu, D., *A Recursive Algorithm for GMDH*, Systems Analysis, Modelling, Simulation, 1990, Vol. 7, No. 7, 533.

Wetherill, G. B., Duncombe, P., Kenward, M., Kallerstrom, J., Paul, S. R. and Vowden, B. J., Regression Analysis with Applications, London: Chapman and Hall, 1986.

Wiener, *What is Information Theory?*, IRE Transactions on Information Theory, 1956, IT-2, 48.

Wilkinson, L., *SYSTAT: The System for Statistics*, Evanston: SYSTAT, 1989.

Yurachkovskiy, Y. P., *Structural Modelling by Observation Sets*, Soviet Automation Control, 1982, 15.

APPENDICES

APPENDIX A

DATA SETS FOR NUMERICAL EXAMPLES

I. Data from Study on Demand of Brand of Liquid Laundry Detergent (Bowerman et al., 1986)

y = demand (in hundreds of thousands) for a bottle of the detergent in a sales period;

x_1 = price (in dollars) of detergent as offered by the manufacturer in the sales period;

x_2 = avg. industry price (in dollars) of competitors' similar detergents in the sales period;

x_3 = advertising expenditure (in hundreds of thousands of dollars) to promote the detergent in the sales period;

x_4 = observed price difference (in dollars) between manufacturer's priced and average industry price (i.e., $x_4 = x_2 - x_1$).

TABLE A.1 LIQUID DETERGENT DEMAND DATA

Sales Period	y	x_1	x_2	x_3	x_4
1	7.38	3.85	3.80	5.50	-0.05
2	8.51	3.75	4.00	6.75	0.25
3	9.52	3.70	4.30	7.25	0.60
4	7.50	3.70	3.70	5.50	0.00
5	9.33	3.60	3.85	7.00	0.25
6	8.28	3.60	3.80	6.50	0.20
7	8.75	3.60	3.75	6.75	0.15
8	7.87	3.80	3.85	5.25	0.05
9	7.10	3.80	3.65	5.25	-0.15
10	8.00	3.85	4.00	6.00	0.15
11	7.89	3.90	4.10	6.50	0.20
12	8.15	3.90	4.00	6.25	0.10
13	9.10	3.70	4.10	7.00	0.40
14	8.86	3.75	4.20	6.90	0.45
15	8.90	3.75	4.10	6.80	0.35
16	8.87	3.80	4.10	6.80	0.30
17	9.26	3.70	4.20	7.10	0.50
18	9.00	3.80	4.30	7.00	0.50
19	8.75	3.70	4.10	6.80	0.40
20	7.95	3.80	3.75	6.50	-0.05
21	7.65	3.80	3.75	6.25	-0.05
22	7.27	3.75	3.65	6.00	-0.10
23	8.00	3.70	3.90	6.50	0.20
24	8.50	3.55	3.65	7.00	0.10
25	8.75	3.60	4.10	6.80	0.50
26	9.21	3.65	4.25	6.80	0.60
27	8.27	3.70	3.65	6.50	-0.05
28	7.67	3.75	3.75	5.75	0.00
29	7.93	3.80	3.85	5.80	0.05
30	9.26	3.70	4.25	6.80	0.55

II. Data from Experiments on Degassing an Aqueous Liquid Stream--The Thackray Data (Wetherill et al., 1986)

y = outlet gas concentration(ppm);

x_1 = vessel pressure;

x_2 = nozzle pressure;

x_3 = flow rate (m^3hr^{-1});

x_4 = inlet gas concentration (ppm);

x_5 = liquid temperature (^oC);

x_6 = pH;

x_7 = liquid concentration.

TABLE A.2 DATA ON DEGASSING AN AQUEOUS LIQUID STREAM

Observation	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	234	0.99	6.0	12.7	497	70	2.0	16.2
2	191	0.99	5.0	12.5	468	71	2.0	16.1
3	234	0.99	3.0	11.2	497	68	2.0	15.4
4	234	0.43	6.2	12.7	582	69	1.8	15.5
5	235	0.43	5.0	12.5	617	68	1.8	15.5
6	191	0.41	4.0	12.0	580	69	1.8	15.6
7	85	0.40	2.9	11.0	475	68	1.8	15.5
8	85	0.32	2.0	9.9	497	67	1.8	15.5
9	85	0.30	2.0	9.9	624	66	1.8	16.4
10	149	0.99	6.0	12.7	504	68	1.8	16.5
11	184	0.99	5.0	12.5	416	69	1.8	16.5
12	161	0.99	3.9	11.9	444	75	1.8	15.9
13	150	0.99	3.0	11.2	475	74	1.8	16.0
14	156	0.99	1.0	8.0	504	71	1.8	15.9
15	89	0.38	2.9	11.0	427	75	1.8	15.9
16	92	0.37	2.5	10.5	355	75	1.8	15.9
17	106	0.33	1.0	8.0	532	70	1.8	15.8
18	149	0.99	6.0	12.7	497	70	1.8	15.9
19	177	0.99	4.0	12.0	497	69	1.8	15.8
20	134	0.45	3.0	11.2	419	71	1.8	15.7
21	92	0.34	3.0	11.2	458	68	1.8	15.7
22	170	0.99	6.0	12.7	451	71	1.8	15.6
23	156	0.99	6.0	12.7	490	71	1.8	15.6
24	227	0.46	1.6	9.2	568	67	1.8	14.8
25	177	0.99	3.0	11.2	454	68	1.8	14.6
26	142	0.99	2.5	10.5	355	74	1.8	15.2
27	64	0.39	3.0	11.2	340	73	1.8	15.1
28	57	0.39	2.5	10.5	369	73	1.8	15.1
29	50	0.38	1.0	8.0	410	73	1.8	15.2
30	206	0.99	1.0	8.0	639	67	1.5	14.9
31	185	0.99	1.5	9.0	589	66	1.5	14.9
32	234	0.99	2.0	9.9	625	66	1.5	14.8
33	128	0.34	1.0	8.0	582	66	1.5	15.3
34	163	0.35	1.4	9.0	596	66	1.5	15.2
35	184	0.35	2.0	9.9	532	66	1.5	15.2
36	199	0.99	3.0	11.2	575	66	1.8	15.3
37	209	0.99	3.0	11.2	582	66	1.8	15.4
38	220	0.99	3.0	11.2	624	66	1.8	15.3
39	205	0.99	3.0	11.2	532	64	1.8	15.3
40	149	0.36	3.0	11.2	639	65	1.8	15.4
41	135	0.36	3.0	11.2	653	66	1.8	15.3
42	135	0.35	3.0	11.2	639	66	1.8	15.4
43	138	0.36	3.0	11.2	646	66	1.8	15.4
44	127	0.39	6.0	12.7	397	75	1.8	15.5

TABLE A.2 (Continued)

Observation	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
45	92	0.38	6.0	12.7	424	75	1.8	15.5
46	85	0.38	6.0	12.7	446	75	1.8	15.6
47	85	0.39	6.0	12.7	424	74	1.8	15.6
48	126	0.41	5.0	12.5	550	75	1.8	16.2
49	57	0.40	5.0	12.5	448	76	1.8	16.2
50	126	0.99	6.0	12.7	416	76	1.8	16.1
51	119	0.99	5.0	12.5	416	76	1.8	16.1
52	142	0.47	6.0	12.7	482	72	1.8	15.7
53	128	0.47	6.0	12.7	482	72	1.8	15.7
54	113	0.47	5.0	12.5	523	72	1.8	15.8
55	134	0.44	4.0	12.0	423	72	1.8	15.7
56	129	0.99	4.0	12.0	404	74	1.8	16.8
57	134	0.99	3.0	11.2	404	74	1.8	16.9
58	134	0.99	2.0	9.9	389	74	1.8	16.9
59	64	0.43	4.0	12.0	404	74	1.8	16.9

APPENDIX B

COMPUTATIONAL MODULES IN MATLAB© FOR THE GMDH ALGORITHM

This appendix gives the major computer programs of the GMDH Algorithm. The programs were developed in Matlab©, version 4.0. These programs were debugged and remodified by making numerous runs on the Wakely (1949) data and the detergent data in Appendix A.

I. Main Matlab© Module for the Original GMDH Algorithm

```
% GMDHorig is a 'm' file to compute the Ivakhnenko polynomials according to
% the original algorithm as developed by Ivankhenko
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by    Cheryl Hild
%               Department of Statistics
%               The University of Tennessee
%               Knoxville, TN  37996
%
% Copyright 4/26/98
% All rights reserved
%
% The output from 'GMDHorig' is saved on your 'orig.out' file.
%
diary orig.out
clear variables
echo off
format long
```

```
format compact
digits(10)
%
% Input your x matrix where the first nt rows are the observations in the training set
%  and the last (n-nt) rows are the checking set.
%
% y is the response vector with the first nt observations being
%  the responses for the training set (1 by n)
%
%
% The matrix of independent variables is:
x=[3.70 4.30 7.25 .6
   3.60 3.80 6.5 .2
   3.80 3.65 5.25 -.15
   3.9 4.1 6.5 .2
   3.7 4.1 7.0 .4
   3.8 4.1 6.8 .3
   3.80 4.3 7 .5
   3.7 4.1 6.8 .4
   3.8 3.75 6.25 -.05
   3.75 3.65 6 -.1
   3.55 3.65 7 .1
   3.6 4.1 6.8 .5
   3.75 3.75 5.75 0
   3.8 3.85 5.8 .05
   3.85 3.8 5.5 -.05
   3.75 4 6.75 .25
   3.7 3.7 5.5 0
   3.6 3.85 7 .25
   3.6 3.75 6.75 .15
   3.80 3.85 5.25 .05
   3.85 4 6 .15
   3.9 4 6.25 .1
   3.75 4.2 6.9 .45
   3.75 4.1 6.8 .35
   3.7 4.2 7.1 .5
   3.8 3.75 6.5 -.05
   3.7 3.9 6.5 .2
   3.65 4.25 6.8 .6
   3.7 3.65 6.5 -.05
   3.7 4.25 6.8 .55];

% The dependent variable is:
y=[9.52 8.28 7.1 7.89 9.1 8.87 9 8.75 7.65 7.27 8.5 8.75 7.67 7.93 ...
   7.38 8.51 7.5 9.33 8.75 7.87 8 8.15 8.86 8.9 9.26 7.95 8 9.21 8.27 9.26];

y=y.';
[n,m]=size(x)
xid=1:m;
xid=xid.';
overallRMIN=100;
```

```
ntrain=input('How many observations are in your training set?')
ncheck=n-ntrain
noiter=1;
maxiter=input('What is the maximum number of generations desired?')
disp('NOTE:  If the RMIN value increases from one generation to the next,')
disp('before the maximum number of generations is reached, the algorithm will stop.')
disp(' ')
Rcutoff=input('What is cutoff value for the regularity criterion?')
Beta=[];
Regs=[];
xback=[];
Coeffpassed=[];
nomodels=m*(m-1)/2
modid=1:nomodels;
modid=modid.';
mk=nomodels;
modpass=zeros(nomodels,maxiter);
for i=1:maxiter
    modpass(:,i)=modid;
end
disp('')
disp('Note:  The Parameter Estimates are Displayed at Each')
disp('       Level in 6x1 vectors of the form (bo,b1,b2,b3,b4,b5,b6).')
disp(' ')
while noiter<=maxiter
   disp(['LEVEL ',num2str(noiter),' OF THE GMDH ALGORITHM'])
[B,R,data,Bback,overallRMIN,noiter,modpass,mk]=ivanorig(m,n,ntrain,ncheck,noite
r,maxiter,x,xback,y,modpass,Rcutoff,overallRMIN);
   if mk-1 == 1
            break
    end
    disp('')
   disp('The Estimated Parameter Coefficients Are:')
    nomodels=mk*(mk-1)/2
    for l=1:nomodels
     disp(['    Model ',num2str(l),':'])
     disp('    _________')
     disp(B(:,l))
   end
    disp('')
   disp('The Values of the Regularity Criterion Are:')
    disp('')
     disp('        Regularity    ')
   disp('--------------------')
   for l=1:nomodels
    disp(['Model ',num2str(l),':  ',num2str(R(l))]);
            disp(' ')
   end
    disp('')
   Beta=[Beta,B];
    Coeffpassed=[Coeffpassed,Bback];
```

```
    if mk<m
            for c=mk:m
                    R=[R;0];
            end
    end
   x=data;
   noiter=noiter+1;
end
noiter=noiter-1;
back = 'y';
back=input('To run backtracing, enter y or Y; otherwise, enter n or N','s');
if back == 'y'|back=='Y'
    [estmodel1]=backorig(m,noiter,nomodels,Beta,Regs,Coeffpassed);
else
    break
end
diary off
```

Ia. Matlab© Function to Estimate Parameters and Compute Values for Regularity Criterion for the Original GMDH Algorithm

```
% The 'ivanorig' function composes and estimates the models at each level of the
algorithm
% based on the data in the training set.
%
% The function then computes the regularity criterion for each of the composed
% models based
% on the observations in the checking set.
%
% Developed by Dr. H. Bozdogan and Cheryl Hild
%
% Written by Cheryl Hild
%            Management Science Program
%            The University of Tennessee
%            Knoxville, TN 37996
%
% Copyright 4/26/98
% All Rights Reserved
%
function[level1,Reg,data,Bback,overallRMIN,noiter,modpass,mk]=ivanorig(m,n,ntrai
n,ncheck,noiter,maxiter,x,xback,y,modpass,Rcutoff,overallRMIN)
x;
[n,m]=size(x);
mk=m;
xtrain=x(1:ntrain,:);
ytrain=y(1:ntrain);
xsqtrain=[];
xsq=[];
for i=1:m
 xvtrain=xtrain(:,i);
 xv=x(:,i);
 for k=1:ntrain
   sqtrain(k)=xvtrain(k)*xvtrain(k);
 end
 for l=1:n
   sq(l)=xv(l)*xv(l);
 end
 xsqtrain=[xsqtrain;sqtrain];
 xsq=[xsq;sq];
end
ztrain=rot90(xsqtrain);
z=rot90(xsq);
xsqtrain=flipud(ztrain);
xsq=flipud(z);
level1=[];
Bback=[];
```

```
ysq=[ones(1,n)];
ysq=ysq.';
Reg=[];
data=[];
p=0;
for i=1:m-1
 for j=2:m
  if i<j&i~j;
    p=p+1;
    model=[i,j];
    x1train=xtrain(:,i);
     x1=x(:,i);
    x2train=xtrain(:,j);
     x2=x(:,j);
     clear xxtrain;
     clear xx;
    for k=1:ntrain;
     xxtrain(k)=x1train(k)*x2train(k);
    end
     for l=1:n;
            xx(l)=x1(l)*x2(l);
     end
    xxtrain=xxtrain.';
     xx=xx.';
    constanttrain=[ones(1,ntrain)];
    constanttrain=constanttrain.';
     constant=[ones(1,n)];
    constant=constant.';
```

ESTIMATION OF PARAMETERS IN QUADRATIC POLYNOMIALS FOR EACH PAIR OF VARIABLES (USING ONLY TRAINING SET)

```
    xrtrain=[constanttrain,x1train,x2train,xsqtrain(:,i),xsqtrain(:,j),xxtrain];
     xr=[constant,x1,x2,xsq(:,i),xsq(:,j),xx];
    design=inv(xrtrain'*xrtrain);
    coeff=design*xrtrain'*ytrain;
     [d,e]=size(coeff);
    yhattrain=xrtrain*coeff;
     yhat=xr*coeff;
    residtrain=ytrain-yhattrain;
     sum(residtrain);
     resid=y-yhat;
     for c=1:n
        residsq(c)=resid(c)*resid(c);
     end
```

```
EVALUATION OF ESTIMATED MODELS: COMPUTATION OF
R^2 AND REGULARITY CRITERION (BASED ON CHECKING SET)

    rss=resid'*resid;
     ybar=sum(y)/n;
    ybarsq=ybar*ybar;
     for c=1:n
       ysq(c)=y(c)*y(c);
     end
     ssy=y'*y;
    rsqr=1-(rss/(ssy-n*ybarsq));
    sighatsq=(resid'*resid)/n;
     sumresidsqck=0;
     sumysqck=0;
     for c=ntrain+1:n
            sumresidsqck=sumresidsqck+residsq(c);
            sumysqck=sumysqck+ysq(c);
     end
     Regcrit(i,j)=sqrt(sumresidsqck/sumysqck);
     Reg=[Reg;Regcrit(i,j)];

CALL FUNCTION TO SCREEN VARIABLES BASED ON VALUES
OF THE REGULARITY CRITERION

     [coeffpass,modpass]=varelimorig(p,coeff,Reg,Rcutoff,modpass,noiter);
     level1=[level1,coeff];
     Bback=[Bback,coeffpass];
    [f,nopoly]=size(level1);
    clear f

CALL FUNCTION TO CREATE MATRIX OF NEW VARIABLES FOR
NEXT LEVEL OF THE ALGORITHM

if noiter<maxiter
    [data,databack]=newdataorig(n,coeff,coeffpass,level1,i,j,p,x,xsq,xx,data,databack);
end
            if p==3
            currentRMIN=min(Reg)
            if currentRMIN<overallRMIN
                    overallRMIN=currentRMIN
            else
                    noiter=maxiter+1
            end
      end
   end
  end
end
```

Ib. Matlab© Function to Screen Variables for Next Level of Algorithm

```
% The 'varelimorig' function compares the obtained value for the regularity criterion
% to the cutoff value provided by the user.
%       For those regularity values which are less than the cutoff value, the associated
%       models are not passed into the next level of the algorithm.
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by  Cheryl Hild
%           Management Science Program
%           The University of Tennessee
%                   Knoxville, TN  37996
%
% Copyright 4/26/98
% All Rights Reserved
%
function[coeffpass,modpass]=varelimorig(p,coeff,Reg,Rcutoff,modpass,noiter)

if Reg(p)>Rcutoff
        modpass(p,noiter)=0;
        for i=1:6
            coeffpass(i)=0;
        end
end
end
```

Ic. Matlab© Function to Compute the Polynomials at all n Data Points and to Create New Matrix of Variables

```
% The 'newdataorig' function creates matrix of new observations by evaluating
% polynomials
% at level j at all n data points and using these values as the x matrix for the next
% generation.
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by  Cheryl Hild
%            Management Science Program
%            The University of Tennessee
%            Knoxville, TN  37996
%
% Copyright 4/26/98
% All Rights Reserved
%
function[data,xback]=newdataorig(n,coeff,coeffpass,level1,i,j,p,x,xsq,xx,data,xback)
v1=[];
v2=[];
v=[zeros(size(n))];
clear Bo
for l=1:n
  Bo(l)=coeffpass(1);
end
Bo=Bo.';
v1=Bo+coeffpass(2)*x(:,i)+coeffpass(3)*x(:,j)+coeffpass(4)*xsq(:,i);
v2=coeffpass(5)*xsq(:,j)+coeffpass(6)*xx;
v=v1+v2;
xback=[xback,v];
if v~=0
    data=[data,v];
end

end
```

II. Main Matlab© Module for the GMDH Algorithm with Information-Based Model Evaluation Criteria

```
% GMDH is an 'm' file to develop the functional form of a model and to
% estimate the parameters through a GMDH type algorithm using information-
% based model evaluation criteria to screen variables at each level.
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by    Cheryl Hild
%               Management Science Program
%               The University of Tennessee
%               Knoxville, TN  37996
% Copyright 4/26/98
% All Rights Reserved
%
% The output from 'GMDH' is saved on your 'GMDH.out' file.
%
diary GMDH.out
clear variables
echo off
format short
digits(10)
%
% Input a (1 x n) row vector of the values for each of the predictor variables
% y is the response (1 x n) vector
%
% The matrix of independent variables is:
x1=[2.80 2.30 2.8 2.73 1.05 2.13 1.8 .8 1.91 2.16 1.77 1.51 1.84 2.03 1.76 1.55
2.73 2.56 1.98 .8 1.8 1.91 1.91 2.13 2.05 .59 .76];
x1=x1.';
x2=[.07 .02 .07 .07 .02 .47 .02 .02 .47 .07 .02 .47 .07 .47 .07 .07 .07 .07 .02 .02
.02 .47 .47 .47 .47 .02 .47];
x2=x2.';
x3=[8.7 8.2 7.4 8.9 7.0 6.9 6.5 6.7 6.9 8.8 7.6 7.5 8.7 7.6 7.8 8.9 7.2 8.4 7.6 6.2
7.3 8.3 8.2 7.6 7.6 6.5 7.4];
x3=x3.';
x=[x1 x2 x3];
% The dependent variable is:
y=[100.3 77.1 102.8 101 88.4 59.4 92 96.2 62.1 98.9 82.4 58.7 ...
   93.7 74 110.4 102.8 108.4 100.7 96.6 ...
   99 75.3 65.7 56.8 53.2 58 99.4 61];
y=y.';
[n,m]=size(x);
xid=1:m;
xid=xid.';
Rcutoff=0;
noiter=1;
```

```
The user inputs the maximum number of levels that the GMDH
will generate and chooses the Information-Based Criteria to be used
maxiter=input('What is the maximum number of generations desired?')
disp(' ')
disp('The following information-theoretic model selection criteria are available for
use:')
disp('AIC-1')
disp('ICOMP-2')
disp('ICOMP(IFIM)-3')
critselect=input('Input the number of the criterion you desire to use.')
disp(' ')
Beta=[];
Regs=[];
xback=[];
mods=[];
Coeffpassed=[];
refno=[];
locator=0;
nomodels=m*(m-1)/2;
mk=nomodels;
modpass=zeros(nomodels,maxiter);
disp('')
disp('Note:  The Parameter Estimates are Displayed at Each')
disp('       Level in 6x1 vectors of the form (bo,b1,b2,b3,b4,b5,b6).')
disp(' ')
The module calls the ivakIC function to compose, evaluate, and screen the
polynomials at each generation of the algorithm
while noiter<=maxiter
   disp(['LEVEL ',num2str(noiter),' OF THE GMDH ALGORITHM'])
[B,data,Bback,noiter,mods,mk,AIC,ICOMP,ICOMPif,locator,refno]=ivakIC(noiter,m
axiter,x,xback,y,mods,critselect,Rcutoff,locator,refno);
   Beta=[Beta,B];
    Coeffpassed=[Coeffpassed,Bback];
    if critselect==1
            Regs=[Regs;AIC];
    elseif critselect==2
            Regs=[Regs;ICOMP];
    elseif critselect==3
            Regs=[Regs;ICOMPif];
    end
   x=data;
   noiter=noiter+1;
   if mk-1 == 1
    break
    end
    disp('')
   disp('The Estimated Parameter Coefficients Are:')
```

```
    nomodels=mk*(mk-1)/2
    for l=1:nomodels
     disp(['     Model ',num2str(l),':'])
     disp('     _________')
     disp(B(:,l))
   end
    disp('')
   disp('The Values of the Information Theoretic Criterion Are:')
    disp('')
       disp('')
    if critselect==1
            disp('          AIC     ')
    disp('--------------------')
    for l=1:nomodels
    disp(['Model ',num2str(l),':  ',num2str(AIC(l))]);
    end
    elseif critselect==2
            disp('          ICOMP     ')
    disp('----------------------')
    for l=1:nomodels
    disp(['Model ',num2str(l),':  ',num2str(ICOMP(l))]);
    end
    elseif critselect==3
            disp('          ICOMP(IFIM)     ')
    disp('---------------------------')
    for l=1:nomodels
    disp(['Model ',num2str(l),':  ',num2str(ICOMPif(l))]);
    end
    end
    disp(' ')
end
noiter=noiter-1;
refno
diary off
back = 'y';
```

'back2' is a function file to backtrace the final model with the minimum value of the chosen criterion. The output of the 'back2' function is the complex model in terms of the original variables.

```
back=input('To run backtracing, enter y or Y; otherwise, enter n or N','s');
if  back=='y'|back=='Y'
    disp('============')
    m;
    noiter;
    [estmodel1]=back2(m,noiter,nomodels,Regs,Coeffpassed,refno,mods);
else
    break
end
```

IIa. Matlab© Function to Estimate Parameters and Compute Values of the Information-Based Criteria for the GMDH Algorithm

```
% The 'ivakIC' function composes and estimates the models at each level of the
algorithm by computing the values of the chosen Information-Based Criterion
%
%
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by  Cheryl Hild
%             Management Science Program
%             The University of Tennessee
%             Knoxville, TN  37996
%
% Copyright 4/26/98
% All Rights Reserved
%
function[level1,data,Bback,noiter,mods,mk,AIC,ICOMP,ICOMPif,locator,refno]=iva
kIC(noiter,maxiter,x,xback,y,mods,critselect,Rcutoff,locator,refno)
%
AIC=[];
ICOMP=[];
ICOMPif=[];
[n,m]=size(x);
mk=m;
xsq=[];
for i=1:m
  xv=x(:,i);
 for l=1:n
    sq(l)=xv(l)*xv(l);
  end
  xsq=[xsq;sq];
end
z=rot90(xsq);
xsq=flipud(z);
level1=[];
Bback=[];
ysq=[ones(1,n)];
ysq=ysq.';
data=[];
p=0;
for i=1:m-1
  for j=2:m
   if i<j&i~j;
     p=p+1;
     model=[i,j];
      mods=[mods; model];
      x1=x(:,i);
```

```
x2=x(:,j);
clear xx;
for l=1:n;
        `xx(l)=x1(l)*x2(l);
end
 xx=xx.';
constant=[ones(1,n)];
constant=constant.';
 xr=[constant,x1,x2,xsq(:,i),xsq(:,j),xx];
design=inv(xr'*xr);
coeff=design*xr'*y;
[d,e]=size(coeff);
yhat=xr*coeff;
 resid=y-yhat;
sum(resid);
for c=1:n
   residsq(c)=resid(c)*resid(c);
end
rss=resid'*resid;
ybar=sum(y)/n;
ybarsq=ybar*ybar;
for c=1:n
  ysq(c)=y(c)*y(c);
end
 ssy=y'*y;
rsqr=1-(rss/(ssy-n*ybarsq));
sighatsq=(resid'*resid)/n;
lackoffit=n*log(2*pi)+n*log(sighatsq)+n;
covBhat=sighatsq*design;
v=diag(covBhat);
w=diag(v);
corrBhat=w^(-1/2)*covBhat*w^(-1/2);
CompCorrBhat=d/2*log(trace(corrBhat)/d)-1/2*log(det(corrBhat));
if critselect==1
        AICcrit(i,j)=lackoffit+2*d;
        AIC=[AIC;AICcrit(i,j)];
elseif critselect==2
        CompCovBhat=(d/2)*log(trace(covBhat)/d)-1/2*log(det(covBhat));
        ICOMPcrit(i,j)=lackoffit+2*CompCovBhat;
        ICOMP=[ICOMP;ICOMPcrit(i,j)];
elseif critselect==3
        a=(d+1)*log((trace(covBhat)+2*sighatsq/n)/d);
        b=-1*log(det(design));
        c=-1*log(2*sighatsq/n);
        ICOMPifcrit(i,j)=lackoffit+2*(a+b+c)
        ICOMPif=[ICOMPif;ICOMPifcrit(i,j)];
else
        disp('Incorrect criterion selection number')
        break;
end
level1=[level1,coeff];
```

```
    if noiter<=maxiter
        [data,databack]=newdata(n,coeff,level1,i,j,p,x,xsq,xx,data,databack);
    end
     if p==mk*(mk-1)/2
    if critselect==1
    [coeffpass,mods,data,locator,refno]=vareliminate(mk,p,level1,AIC,Rcutoff,mods,
          noiter,data,n,locator,refno);
    elseif critselect==2
    [coeffpass,mods,data,locator,refno]=vareliminate(mk,p,level1,ICOMP,Rcutoff,
          mods,noiter,data,n,locator,refno);
    elseif critselect==3
    [coeffpass,mods,data,locator,refno]=vareliminate(mk,p,level1,ICOMPif,Rcutoff,
          mods,noiter,data,n,locator,refno);
    end
    end
    Bback=[Bback,coeffpass];
    [f,nopoly]=size(level1);
     clear f
      end
  end
end
```

IIb. Matlab© Function to Compose New Variables and Create New Matrix of Predictor Variables for Each Generation

```
% The 'newdata' function creates matrix of new observations by evaluating
% polynomials
% at level j at all n data points and using these values as the x matrix for the next
% generation.
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by  Cheryl Hild
%           Management Science Program
%           The University of Tennessee
%           Knoxville, TN  37996
%
% Copyright 4/26/98
% All Rights Reserved
%
function[data,xback]=newdata(n,coeff,level1,i,j,p,x,xsq,xx,data,xback);
v1=[];
v2=[];
v=[zeros(size(n))];
clear Bo
for l=1:n
    Bo(l)=coeff(1);
end
Bo=Bo.';
v1=Bo+coeff(2)*x(:,i)+coeff(3)*x(:,j)+coeff(4)*xsq(:,i);
v2=coeff(5)*xsq(:,j)+coeff(6)*xx;
v=v1+v2;
data=[data,v];
xback=data;
end
```

IIc. Matlab© Function to Screen Variables for Next Level of Algorithm

```
% The 'vareliminate' function compares the obtained value for the information criterion
%  to the maximum value of the criterion in the geneneration.
%  A selection ratio is computed as the difference between the computed value and the
%  maximum value divided by the average difference in the generation.
%
%  Those models with a selection ratio higher than the cutoff value input by the user
%  are screened and not passed to the next level of the algorithm.
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by  Cheryl Hild
%             Management Science Program
%             The University of Tennessee
%             Knoxville, TN  37996
%
% Copyright 4/26/98
% All Rights Reserved
%
function[coeffpass,mods,updatedata,locator,refno]=vareliminate(mk,p,coeff,crit,Rcuto
ff,mods,noiter,data,n,locator,refno)
disp('start of elimination function')
refno=[refno,p];
data;
updatedata=[];
Diff=[];
ration=[];
coeffpass=coeff;
maxcrit=max(crit);
if p~=1
for i=1:p
    Diff(i)=maxcrit-crit(i);
end
avgdiff=sum(Diff)/p;
for i=1:p
    ratio(i)=Diff(i)/avgdiff;
end
ratio;
coeffpass;
updatedata;
```

```
    for i=1:p
        if ratio(i)<Rcutoff
            mods(locator+i,1)=0;
            mods(locator+i,2)=0;
            for j=1:6
                coeffpass(j,i)=0;
            end
        end
        if ratio(i)>=Rcutoff
            updatedata=[updatedata,data(:,i)];
        end
    end
end
updatedata;
mods;
coeffpass
locator=locator+p;
end
end
```

III. Matlab© Function to Backtrace the Model in the Last Generation of the GMDH Algorithm with the Minimum Value of the Criterion

```
% The 'back2'backtraces the final model from the GMDH and expresses it in terms
% of the original predictor variables.
%
% Developed by  Dr. H. Bozdogan and Cheryl Hild
%
% Written by  Cheryl Hild
%          Management Science Program
%          The University of Tennessee
%          Knoxville, TN  37996
%
% Copyright 4/26/98
% All Rights Reserved
%

function[b] = back2(m,noiter,nomodels,Regs,Bback,refno,mods)
diary backtracing.out
models=[];
format short;
modlocator1=1;
modlocator2=0;
[r,nogens]=size(refno);
B=Bback;
nullvector=zeros(1:6);
[rows,cols]=size(B);
disp('Input the number of generation of the algorithm with the model you wish to
backtrace.')
nolevels=input(['This number must be <= ', num2str(nogens)])
if nolevels==1
    modlocator1=1;
    modlocator2=modlocator1+refno(1)-1;
    mods(modlocator1:modlocator2,:);
elseif nolevels>1
    for i=1:nolevels-1
            modlocator1=modlocator1+refno(i);
    end
    modlocator2=modlocator1+refno(i+1)-1;
end
levelRegs=Regs(modlocator1:modlocator2);
minRegs=min(levelRegs)
modelcolumn=find(Regs==minRegs)
y='y';
x='x';
eqform='b(1)+b(2)*x+b(3)*y+b(4)*x^2+b(5)*y^2+b(6)*x*y';
eqform=sym(eqform);
models=sym(modelcolumn,1,'eqform');
```

```
for i=1:modelcolumn
        b=B(:,i);
        b=b.';

    if b~=nullvector
        step1=subs(eqform,num2str(b(1)),'b(1)');
        step2=subs(step1,num2str(b(2)),'b(2)');
        step3=subs(step2,num2str(b(3)),'b(3)');
        step4=subs(step3,num2str(b(4)),'b(4)');
        step5=subs(step4,num2str(b(5)),'b(5)');
        esteq=subs(step5,num2str(b(6)),'b(6)');
        models=sym(models,i,1,esteq);
    end
        k=k+1;
end
models
varform='x(t)';
varform=sym(varform);
index=0;
for i=1:refno(1)
        sym(models);
        y=sym(models,i,1);
        if sym(y(1))~='e'
            var1=subs(varform,num2str(mods(i,1)),'t');
            var2=subs(varform,num2str(mods(i,2)),'t');
            Z=sym(models,i,1);
            A=subs(Z,var1,'x');
            estmod=subs(A,var2,'y');
            models=sym(models,i,1,estmod);
        end
end
models;

disp('LEVEL ONE IN ORIGINAL VARIABLES')

if nolevels==1
    disp('The best approximating model in generation ')
    disp([num2str(nolevels),': '])
    finalmodel=sym(models,modelcolumn,1)
    break
end

currentindex=0;
levelindex=[];
for i=1:nolevels
    currentindex=currentindex+refno(i);
    levelindex(i)=currentindex;
end
levelindex(nolevels)=modelcolumn;
levelindex;
```

```
nodeletes=0;
for i=1:modelcolumn
    if mods(i,1)==0&mods(i,2)==0
            nodeletes=nodeletes+1;
    end
end
finalmodels=sym(modelcolumn-nodeletes,1,'Q');
finalmodid=[];

for j=1:levelindex(nolevels)
    if mods(j,1)~=0&mods(j,2)~=0
            index=index+1;
            finalmodid=[finalmodid;mods(j,:)];
            equate=sym(models,j,1);
            tosub=sym(finalmodels,index,1);
            newequate=subs(tosub,equate,'Q');
            finalmodels=sym(finalmodels,index,1,newequate);
    end
end
finalmodels;
for k=nolevels:-1:2
    for i=levelindex(k):-1:levelindex(k-1)+1
            if mods(i,1)==0&mods(i,2)==0
                    refno(k)=refno(k)-1;
            end
    end
end
for i=1:refno(1)
    if mods(i,1)==0&mods(i,2)==0
            refno(1)=refno(1)-1;
    end
end
refno;
currentindex=0;
for i=1:nolevels
    currentindex=currentindex+refno(i);
    levelindex(i)=currentindex;
end
levelindex(nolevels)=modelcolumn-nodeletes;
modelcolumn=levelindex(nolevels);
levelbreak=[0, levelindex];
finalmodid;

if nolevels>2
for i=2:nolevels-1
    for j=levelindex(i-1)+1:levelindex(i)
            subid1=finalmodid(j,1);
            subid2=finalmodid(j,2);
            sublocator1=levelbreak(i-1)+subid1;
            sublocator2=levelbreak(i-1)+subid2;
```

```
            Maineq=sym(finalmodels,j,1);
            Subeq1=sym(finalmodels,sublocator1,1);
            Subeq2=sym(finalmodels,sublocator2,1);
            Firstsub=subs(Maineq,Subeq1,'x');
            Secondsub=subs(Firstsub,Subeq2,'y');
            q=expand(Secondsub);
            finalmodels=sym(finalmodels,j,1,q);
    end
end
end
digits(7)
subid1=finalmodid(modelcolumn,1);
subid2=finalmodid(modelcolumn,2);
i=nolevels;
sublocator1=levelbreak(i-1)+subid1;
sublocator2=levelbreak(i-1)+subid2;
Maineq=sym(finalmodels,modelcolumn,1)
Subeq1=sym(finalmodels,sublocator1,1);
Subeq1=vpa(Subeq1,6);
Subeq2=sym(finalmodels,sublocator2,1);
Subeq2=vpa(Subeq2,6);
Firstsub=subs(Maineq,Subeq1,'x');
Secondsub=subs(Firstsub,Subeq2,'y');
q=expand(Secondsub)
diary off
```

VITA

Cheryl Hild, a senior statistician for Six Sigma Associates, specializes the application of statistical methodologies in business and manufacturing processes. Prior to joining Six Sigma Associates, Cheryl served as program manager and instructor for The University of Tennessee's Management Development Center. She also served as the Director of the Center for Quality and Productivity at Pellissippi State Community Technical College. She has worked within many organizations implementing statistical process control and design of experiments, including Whirlpool Corporation, AlliedSignal Aerospace Services, and Johnson Controls World Services.

Cheryl is the co-author of *The Power of Statistical Thinking: Improving Industrial Processes*, published by Addison-Wesley in 1995. In 1996, Cheryl received the Chancellor's Citation for Extraordinary Professional Promise at The University of Tennessee. She received her B.S. in Quantitative Methods from The University of Alabama in Birmingham in 1988.