8-2008

Learning Confirmatory Patterns in Exploratory Factor Analysis Using ICOMP and Genetic Algorithm

Hongwei Yang
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Hongwei Yang entitled "Learning Confirmatory Patterns in Exploratory Factor Analysis Using ICOMP and Genetic Algorithm." I have examined the final electronic 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 Education.

Schuyler W. Huck, Hamparsum Bozdogan, Major Professor

We have read this dissertation and recommend its acceptance:

Tricia McClam, Russell Zaretzki

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

We are submitting herewith a dissertation written by Hongwei Yang entitled "Learning Confirmatory Patterns in Exploratory Factor Analysis Using ICOMP and Genetic Algorithm." We have examined the final electronic 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 Education.

Schuyler W. Huck,
Major Professor and Committee Co-chair

Hamparsum Bozdogan,
MS Advisor and Committee Co-chair

We have read this dissertation
and recommend its acceptance:

Tricia McClam,
Committee Member

Russell Zaretzki,
Committee Member

Accepted for the Council:

Carolyn R. Hodges,
Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

LEARNING CONFIRMATORY PATTERNS IN EXPLORATORY FACTOR ANALYSIS USING ICOMP AND GENETIC ALGORITHM

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Hongwei Yang

August 2008

DEDICATION

I dedicate this work to my father Yonghua Yang, my mother Tongzhen Wang, my brother Dong Yang, his wife Rongrong Wang, and their son Rui Yang.

ACKNOWLEDGMENTS

I would like to thank my advisor, Prof. Schuyler W. Huck, for his encouragement, interest, and patience. I would also like to thank Prof. Bozdogan for sharing his knowledge which has enriched my study in Statistics. The work from Prof. Tricia McClam and Prof. Russell Zaretzki is also highly appreciated.

ABSTRACT

The dissertation intends to develop a new approach to the identification of the best factor pattern structure. This new approach is a multivariate regression analysis where factor scores are regressed on original variables. The dissertation shows the versatility of information model selection criteria, Bozdogan's ICOMP-type criteria in particular, in two types of modeling problems: determining the number of factors in factor analysis and working as the fitness function for Genetic Algorithm.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction and Purpose

1.1 Overview

Factor analysis is a widely used multivariate statistical technique the principal objective of which is to attain a parsimonious description of the observed data (Harman, 1976), or to reduce the dimensionality of the observed data. There are two types of factor analysis: *exploratory factor analysis* (EFA) and *confirmatory factor analysis* (CFA).

EFA is often used to explore the possible underlying factor structure of a set of observed variables without imposing a preconceived structure on the outcome (Child, 1990) whereas CFA usually serves to verify the hypothesized factor structure configuration of a set of observed variables. CFA requires specification of a model a priori whereas EFA is only an orderly simplification of interrelated variables (Suhr, n.d.). A major drawback/disadvantage of EFA is its inability to specify correlational relationships in the factor model (Munro, 2004). Because EFA is far more common than CFA in social sciences (Garson, 2008), it is certainly desirable to add a factor pattern identification algorithm to regular EFA

so that this type of analysis leads to a specified factor model solution which provides a satisfactory model-data fit. Therefore, this study intends to present such a technique that finally unifies EFA and CFA.

EFA assumes there are certain common factors and certain specific factors (Lawley & Maxwell, 1971) and, thus, there are usually two problems associated with this type of modeling scenario. First, no prior knowledge is available regarding what is the true dimensionality of the original set of variables. Second, no information is available regarding what the relationship is between extracted factors and observed variables. Given a finite sample, the first problem relates to how many factors should be extracted so that the information contained in the original set of variables is sufficiently explained whereas the second problem involves an investigation of the regression relationship between original variables and estimated factor scores from the factor model solution. This study addresses both problems with the aid of information model selection criteria and Genetic Algorithm.

1.2 Problem One

For the first problem of determining the number of factors in factor analysis, the use of information model selection criteria overcomes problems involved in the traditional goodness-of-fit test approach (Bozdogan & Ramirez, 1987). The traditional approach, proposed by Lawley (1940, 1942), is to use a sequence of χ^2 hypothesis tests for judging the adequacy of the factor model when $m = 1, 2, ..., M$ factors are fitted. These χ^2 tests, for which the null hypothesis is that the current $m - factor$ model fits the data well, are very sensitive to the sample size and a large sample can easily cause a good factor model to be rejected.

In addition to the sample size issue, although the level of significance for each test $\alpha_{(1)}, \alpha_{(2)}, ..., \alpha_{(M)}$ is known, the overall level of significance for the entire model selection procedure is unknown (Anderson & Rubin, 1976). In other words, it is impossible to choose a level of significance probability which accounts for the number of hypothesis tests being performed. As a result, the critical values from the sequence of χ^2 tests are not adjusted from one model to another. And this causes the conclusion on the number of factors from the hypothesis test approach to be problematic due to the inflated Type I error rate.

On the other hand, information model selection criteria are not hypothesis

tests, so no level of significance probability is involved. Stated differently, the level of significance probability is already implicitly incorporated within the model selection criteria which depend on the specific functional form of the penalty component of the criteria (Bozdogan, 2000). These criteria tend to choose the factor model which is the least likely to be rejected to be the best approximating model. In addition, these criteria map how well the factor model fits the data to a scalar value, which makes the comparison of factor models straightforward.

1.3 Problem Two

For the second problem of identifying the factor pattern, the use of Genetic Algorithm provides a highly intelligent and computationally feasible approach to selecting the optimal multivariate regression subset of predictors (Bearse & Bozdogan, 2000). This subset of predictors is a required component for determining the factor pattern. After the number of factors m is determined, the factor model can be obtained and the factor score matrix $\mathbf{F}$ estimated. To investigate how the factor solution relates to the original variables, the number of which is p, involves regressing $\mathbf{F}$ on the original raw data $\mathbf{X}$. The total number of possible subset regression models is 2^{mp}. With even moderate values of m and p, the task of

evaluating all subset models is extremely computationally expensive (Bearse & Bozdogan, 2000).

As a result, a technique is called for which is capable of selecting the best-approximating subset model from a portfolio of competing models in a reasonable amount of time. Proposed in this study is the use of Genetic Algorithm which iteratively compares models and chooses better ones based on the concept of natural selection (Bearse & Bozdogan, 2000). This technique is applicable to both linear and nonlinear modeling problems. Demonstrated in this study is how it is used in multivariate regression analysis under Gaussian errors. An information model selection criterion of choice will be used as the fitness function for Genetic Algorithm.

1.4 Outline of the Dissertation

The outline of the dissertation is as follows. In Section 2, the theory of maximum likelihood factor analysis is briefly covered. Section 3 gives the formulas for six information model selection criteria used to choose the number of factors in exploratory factor analysis. Presented in Section 4 is an introduction of Genetic Algorithm with one type of information criterion as the fitness function, and Sec-

tion 5 is on how MVR model parameters are estimated given a selected subset of predictors. Numerical examples are given in Section 6 that show the application of the technique proposed in the study.

2 Factor Analysis Model & Maximum Likelihood Factor Analysis

2.1 Orthogonal Factor Model

Let $\mathbf{x}$ be a vector of p observed variables x_1, x_2, ..., x_p that satisfies the following condition:

$$\mathbf{x} \tilde{} \mathbf{N}_p \left(\boldsymbol{\mu}, \boldsymbol{\Sigma} \right). \tag{1}$$

It is assumed that the covariance structure of the $\mathbf{x}$ vector $\boldsymbol{\Sigma}$ is of full rank p. It is also assumed that the mean vector $\boldsymbol{\mu} = \mathbf{0}$.

Then the $m - factor\ orthogonal\ factor\ model$ holds for x_1, x_2, ..., x_p in the following form:

$$\underset{(p\times 1)}{\mathbf{x}} = \underset{(p\times m)}{\boldsymbol{\Lambda}} \underset{(m\times 1)}{\mathbf{f}} + \underset{(p\times 1)}{\boldsymbol{\varepsilon}}, \tag{2}$$

where

$\mathbf{x}$ is a p by 1 vector representing the observed data,

$\boldsymbol{\Lambda}$ is a p by m factor loading matrix,

$\mathbf{f}$ is an m by 1 vector representing m common factors,

$\boldsymbol{\varepsilon}$ is a p by 1 vector representing residuals.

That is to say, each of the observed variables x_1, x_2, ..., x_p can be written as a linear combination of the m common factors plus a residual term:

$$\begin{aligned} x_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + ... + \lambda_{1m}f_m + \varepsilon_1, \\ x_2 &= \lambda_{21}f_1 + \lambda_{22}f_2 + ... + \lambda_{2m}f_m + \varepsilon_2, \\ &\text{.....}, \\ x_p &= \lambda_{p1}f_1 + \lambda_{p2}f_2 + ... + \lambda_{pm}f_m + \varepsilon_p, \end{aligned} \tag{3}$$

where the weight matrix, $\underset{(p\times m)}{\boldsymbol{\Lambda}} = [\lambda_{ij}]$, $i = 1, 2, ..., p$, $j = 1, 2, ..., m$, represents factor loadings.

Also, the model in Equation 2 imposes a covariance structure on $\mathbf{x}$ given by

$$Cov(\mathbf{x}) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{4}$$

$$Cov(\mathbf{x}, \mathbf{f}) = \boldsymbol{\Lambda}. \tag{5}$$

In Equation 4, $\underset{(p\times p)}{\boldsymbol{\Psi}} = Diag(\Psi_1, \Psi_2, ..., \Psi_p)$ is a diagonal matrix the elements of which are variances for the residual terms ε_1, ε_2, ..., ε_p in $\underset{(p\times 1)}{\boldsymbol{\varepsilon}}$. $\underset{(p\times p)}{\boldsymbol{\Psi}}$ is the

matrix of specific variances which represents the *uniqueness* of each variable in the observed vector $\underset{(p\times 1)}{\mathbf{x}}$.

The orthogonal factor model is usually based on the following assumptions

$$\mathbf{f} \tilde{} \mathbf{N}_m \left(\mathbf{0}, \mathbf{I}_m\right), \tag{6}$$

where

$$m \leq p, \tag{7}$$

$$\boldsymbol{\varepsilon} \tilde{} \mathbf{N}_p \left(\mathbf{0}, \boldsymbol{\Psi}\right), \tag{8}$$

$$Cov\left(\mathbf{f}, \boldsymbol{\varepsilon}\right) = \mathbf{0}, \tag{9}$$

$$\underset{(m\times p)}{\boldsymbol{\Lambda}'} \underset{(p\times p)}{\boldsymbol{\Psi}^{-1}} \underset{(p\times m)}{\boldsymbol{\Lambda}} = \underset{(m\times m)}{\boldsymbol{\Delta}}, \tag{10}$$

where $\underset{(m\times m)}{\boldsymbol{\Delta}} = Diag\left(\delta_1, \delta_2, \ ..., \delta_m\right)$ with $\delta_1 > \delta_2 > ... > \delta_m$ is called *matrix unique condition* because of the multiplicity of choices for the factor loading matrix $\boldsymbol{\Lambda}$ (Johnson & Wichern, 1982). This constraint is used to force the iterative maximum likelihood estimates to converge to a unique solution.

2.2 Maximum Likelihood Analysis of Factor Models

In maximum likelihood common factor analysis under the orthogonal factor model, the aim is to estimate the factor loading matrix $\underset{(p\times m)}{\mathbf{\Lambda}}$ and the specific variance matrix $\underset{(p\times p)}{\mathbf{\Psi}}$ given the number of factors using the method of maximum likelihood to explain an empirical correlation matrix (Harman, 1976). In other words, the interest is in the MLE of the factor loading matrix $\widehat{\mathbf{\Lambda}}$ and the MLE of the unique variance matrix $\widehat{\mathbf{\Psi}}$ when the *uniqueness constraint* in Equation 10 is satisfied.

In this study, $MATLAB^{\text{TM}}$'s $factoran$ function is used to find $\widehat{\mathbf{\Lambda}}$ of $\mathbf{\Lambda}$ and $\widehat{\mathbf{\Psi}}$ of $\mathbf{\Psi}$ given the number of extracted factors m. This function $factoran$ standardizes the observed data matrix $\underset{(n\times p)}{\mathbf{X}}$ to zero mean and unit variance before estimating $\mathbf{\Lambda}$ and $\mathbf{\Psi}$. As a result, $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{\Psi}}$ are returned in terms of the standardized variables, or $\widehat{\mathbf{\Lambda}} = \widehat{\mathbf{\Lambda}}_z$ and $\widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Psi}}_z$. In other words, $\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Lambda}}' + \widehat{\mathbf{\Psi}} = \widehat{\mathbf{\Lambda}}_z\widehat{\mathbf{\Lambda}}_z' + \widehat{\mathbf{\Psi}}_z$ is an estimate of the correlation matrix of the original data $\underset{(n\times p)}{\mathbf{X}}$ (The MathWorks, Inc., 2007):

$$\underset{(p\times p)}{\widehat{\mathbf{R}}_m} = \underset{(p\times m)}{\widehat{\mathbf{\Lambda}}_z}\,\underset{(m\times p)}{\widehat{\mathbf{\Lambda}}_z'} + \underset{(p\times p)}{\widehat{\mathbf{\Psi}}_z}\,. \tag{11}$$

It should be noted that Equation 4 and Equation 11 are equivalent of each other when the original raw data $\mathbf{X}$ are standardized to zero mean and unit variance due to the scale invariance property of the maximum likelihood estimators. When

$\mathbf{\Sigma}$ has the structure $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$, then the population correlation matrix $\mathbf{P}$ can be factored as

$$\begin{aligned} \mathbf{P} &= \mathbf{V}^{-1/2}\mathbf{\Sigma}\mathbf{V}^{-1/2} \\ &= \left(\mathbf{V}^{-1/2}\mathbf{\Lambda}\right)\left(\mathbf{V}^{-1/2}\mathbf{\Lambda}\right)' + \mathbf{V}^{-1/2}\mathbf{\Psi}\mathbf{V}^{-1/2} \\ &= \mathbf{\Lambda}_z\mathbf{\Lambda}_z' + \mathbf{\Psi}_z, \end{aligned} \tag{12}$$

where

$$\begin{aligned} \mathbf{V}^{-1/2} &= Diag\left(\sigma_{11}^{-1/2}, \sigma_{22}^{-1/2}, ..., \sigma_{pp}^{-1/2}\right) \\ &= Diag\left(\left(\sigma_1^2\right)^{-1/2}, \left(\sigma_2^2\right)^{-1/2}, ..., \left(\sigma_p^2\right)^{-1/2}\right). \end{aligned} \tag{13}$$

Therefore

$$\widehat{\mathbf{P}} = \widehat{\mathbf{R}}_m = \widehat{\mathbf{\Lambda}}_z\widehat{\mathbf{\Lambda}}_z' + \widehat{\mathbf{\Psi}}_z, \tag{14}$$

which can be used equivalently because of the powerful invariance property of the maximum likelihood estimators. In addition, extracting the factors from the sample correlation matrix rather than the sample covariance matrix can usu-

ally achieve much faster convergence in the optimization process (Bozdogan & Ramirez, 1987).

3 Determining the Number of Factors Using Information Criteria

3.1 A Traditional Approach

In most cases when factor analysis is performed, there is no information regarding what the true underlying dimension m is. As a result, the goal is to find a best approximating factor model based on the finite set of available data. To that end, determination of an appropriate number of factors is a task that has to be accomplished.

In the literature, one traditional approach to selecting the number of factors is the classical goodness-of-fit test which has been under criticism for not adjusting the critical value from one model to another when a set of hypotheses are being tested (Everitt, 1984; Lawley & Maxwell, 1971). In practice, the problem of determining the number of factors is usually not testing just one hypothesis, but rather involves multiple test decisions (Anderson & Rubin, 1956). However, the overall level of significance for fitting $m = 1, 2, ..., M$ factors is unknown.

Proposed in this study is a new approach to determining the number of factors in factor analysis. This new technique involves the use of Information Complexity

Model Selection Criteria, or ICOMP (Bearse & Bozdogan, 2000; Bensmail & Bozdogan, 2002; Bozdogan, 1996, 2000; Bozdogan & Haughton, 1998). ICOMP-type criteria, like AIC (Akaike, 1973, 1987), CAIC (Bozdogan & Ramirez, 1987) and SBC (Schwartz, 1978), belong to a larger family of entropy-based model selection criteria. Although the use of AIC, CAIC and SBC in factor analysis is well documented in the literature (Akaike, 1978; Bozdogan & Ramirez, 1987; Lopes & West, 2004), the research on applying ICOMP-type criteria to such modeling problems is sparse. One of only the few studies that use ICOMP-type criteria in factor analysis is the one by Bozdogan and Shigemasu (1998). In addition, Liu and Bozdogan (2004) have applied ICOMP criteria to principal component analysis (PCA), another statistical technique the major goal of which is similar to that of factor analysis.

3.2 Mathematical Forms of Information Criteria

Here, the search for the optimal number of factors is based on 3 ICOMP criteria: $ICOMPIFIM$, $ICOMPIFIM_{PEU\_Mis}$ and $ICOMPIFIM_{PEU\_Mis\_LN}$. In addition, AIC, $CAIC$ and SBC are also used for the purpose of comparing model selection results. A total of six model selection criteria are to be scored using the

finite original data for each candidate factor model. Based on one criterion, the optimal number of factors is selected as the one that leads to its minimization.

Without providing detailed mathematical proof, analytical forms of the six information criteria for the orthogonal factor model are given below. These formulas use standard outputs from $factoran$.

$$
\begin{aligned}
ICOMPIFIM &= -2\log L\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Lambda}},\widehat{\boldsymbol{\Psi}}\right)+2C_1\left(\widehat{F}^{-1}\right), && (15)\\
ICOMPIFIM_{PEU\_Mis} &= -2\log L\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Lambda}},\widehat{\boldsymbol{\Psi}}\right)+2\left(\frac{ns}{n-s-2}\right) && (16)\\
&\quad +2C_1\left(\widehat{F}^{-1}\right), \\
ICOMPIFIM_{PEU\_Mis\_LN} &= -2\log L\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Lambda}},\widehat{\boldsymbol{\Psi}}\right)+2\left(\frac{ns}{n-s-2}\right) && (17)\\
&\quad +\left[\log\left(n\right)\right]C_1\left(\widehat{F}^{-1}\right), \\
AIC &= -2\log L\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Lambda}},\widehat{\boldsymbol{\Psi}}\right)+2k, && (18)\\
CAIC &= -2\log L\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Lambda}},\widehat{\boldsymbol{\Psi}}\right)+\left[\log\left(n\right)+1\right]k, && (19)\\
SBC &= -2\log L\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Lambda}},\widehat{\boldsymbol{\Psi}}\right)+\left[\log\left(n\right)\right]k, && (20)
\end{aligned}
$$

where

$$-2\log L\left(\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Lambda}},\widehat{\boldsymbol{\Psi}}\right)=np\log\left(2\pi\right)+n\log\left|\widehat{\mathbf{R}}_m\right|+ntr\left(\widehat{\mathbf{R}}_m^{-1}\mathbf{R}\right), \tag{21}$$

$$\widehat{\mathbf{R}}_m=\widehat{\boldsymbol{\Lambda}}_z\widehat{\boldsymbol{\Lambda}}_z'+\widehat{\boldsymbol{\Psi}}_z, \tag{22}$$

$$2C_1\left(\widehat{F}^{-1}\right)=s\log\left[\frac{\left(tr\widehat{\mathbf{R}}_m\right)tr\left(\mathbf{F}'\mathbf{F}\right)^{-1}+\frac{1}{2n}\left(tr\widehat{\mathbf{R}}_m^2+\left(tr\widehat{\mathbf{R}}_m\right)^2+2\sum_{j=1}^{p}\left(r_{jj}\right)^2\right)}{s}\right] \tag{23}$$

$$-\left(p+m+1\right)\log\left|\widehat{\mathbf{R}}_m\right|+p\log\left|\mathbf{F}'\mathbf{F}\right|+\frac{1}{2}p\left(p+1\right)\log\left(n\right)-p\log\left(2\right),$$

$$s=\dim\left(\widehat{F}^{-1}\right)=rank\left(\widehat{F}^{-1}\right)=\frac{1}{2}\left[2pm+p\left(p+1\right)\right], \tag{24}$$

$$k=\left(mp+p\right)-\frac{1}{2}m\left(m-1\right), \tag{25}$$

and

$\widehat{\mathbf{R}}_m$: Estimated factor model correlation matrix,

$\mathbf{R}$: Sample correlation matrix,

$\mathbf{F}$: Estimated factor score matrix,

$\widehat{F}^{-1}$: Estimated inverse-Fisher information matrix,

s : Dimension of $\widehat{F}^{-1}$,

$\mathbf{C}_1\left(\widehat{F}^{-1}\right)$: Complexity of $\widehat{F}^{-1}$,

$-2\log L\left(\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Lambda}}, \widehat{\boldsymbol{\Psi}}\right)$: Minus twice maximized log likelihood function,

n : Sample size,

p : Number of variables in the original data set,

m : Number of common factors,

k : Number of free parameters in the factor model.

For each candidate factor model, a total of six information criteria are scored. And candidate factor models are compared using each criterion, respectively. For a finite original data set of p variables, $factoran$ has an upper limit M on the number of factors that can be extracted.

When m assumes an integer value that falls into the interval determined by

1 and M, the goal is to find the factor model that achieves the minimum on $ICOMPIFIM$, the information criterion of choice, and this model is selected as the best approximating model for the data set. This factor model also serves as the basis for all follow-up analyses.

Next, the $(n \times p)$ original data $\mathbf{X}$ is converted to the $(n \times m)$ factor score matrix $\mathbf{F}$, then the factor structure being sought can be determined by regressing $\mathbf{F}$ on $\mathbf{X}$. That is to say, the regression coefficient matrix $\mathbf{B}$ in $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ is to be estimated. When $m > 1$, this is a *multivariate regression* (MVR) analysis where there are m response variables and p predictors. As a result, an optimal MVR subset of predictors is to be identified which is reasonably parsimonious and achieves pretty good predictive power. This is where *Genetic Algorithm* comes into play.

4 Genetic Algorithm

4.1 From Factor Analysis to Multivariate Regression

When the number of factors is already determined using information criteria, the factor solution is obtained. And the factor score matrix $\underset{(n\times m)}{\mathbf{F}}$ is estimated using Bartlett's Weighted Least Squares method. Given $\underset{(n\times m)}{\widehat{\mathbf{F}}}$ and the original data $\underset{(n\times p)}{\mathbf{X}}$, the regression coefficient matrix $\mathbf{B}$ in $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ is to be estimated. When two or more factors are extracted or $m \geq 2$, the problem of estimating $\mathbf{B}$ becomes that of multivariate regression modeling because there is more than one response variable in the estimated factor score matrix $\widehat{\mathbf{F}}$ and these response variables have to be considered simultaneously.

Before estimating the coefficient matrix $\mathbf{B}$, it is necessary to identify an MVR subset matrix that shows to which group of predictors each response variable is related so that the fitted regression model reaches a satisfactory level of predictive power. This subset of predictors matrix is the basis of a two-step approach to estimating $\mathbf{B}$ which is to be used in a later section of the study. In other words, a subset matrix consisting of only 1's and 0's is being sought here in this section that

has the same number of rows and columns as the regression coefficient matrix **B**. Each 1 in the matrix indicates the inclusion of a predictor for a response variable whereas each 0 the opposite. To obtain such a subset matrix, Genetic Algorithm (GA) is used.

4.2 Genetic Algorithm with ICOMP as the Fitness Function

Genetic Algorithm is a stochastic search method that is widely used in model selection problems from a wide variety of fields such as Engineering, Economics, Game Theory, Biology etc. (Holland, 1975). As a clever non-local optimization algorithm, GA is capable of pruning combinatorially large numbers of sub-models to obtain an optimal or near-optimal MVR subset (Bearse & Bozdogan, 2000) so the algorithm is effective in terms of solving problems where a large number of possible solutions exist.

GA is based on evolution and natural selection concepts in Biology and uses a series of genetic operators in its implementation, such as crossover & mutation. And it selects a champion model by maximizing or minimizing a fitness function mapping the performance of a candidate model to a scalar value with which a comparison of competing models becomes easy and straightforward.

GA represents a model using a binary coding, called *chromosome*, with 1 indicating the presence of a variable in the model and 0 the absence. Therefore, each subset MVR model is represented by a binary string on which a 1 includes the predictor and a 0 excludes it. The length of the GA binary string is equal to the number of predictors, p in $\underset{(n\times p)}{\mathbf{X}}$, times the number of responses, m in $\underset{(n\times m)}{\widehat{\mathbf{F}}}$.

Suppose an MVR model is to be fitted where two responses f_1 and f_2 and three predictors x_1, x_2 and x_3 are taken into account.

$$\begin{aligned} f_{i1} &= x_{i1}B_{11} + x_{i2}B_{21} + x_{i3}B_{31}, \\ f_{i2} &= x_{i1}B_{12} + x_{i2}B_{22} + x_{i3}B_{32}, \\ i &= 1, 2, ..., n. \end{aligned} \tag{26}$$

A binary string of 110010 generated by GA indicates that, in the MVR model, x_1 and x_2 are selected for f_1 whereas x_2 is the only predictor selected for f_2.

GA also needs an objective function or a fitness function on which to base the decision of model choice. The fitness function used in the GA maps the performance of a candidate model to a scalar value. The model that maximizes or minimizes the fitness function is selected as the champion model. Although the choice of a GA fitness function is plenty, this study implements Bozdogan's

Information Complexity Criterion (ICOMPIFIM) which penalizes parameter interactions/redundancy as well as the number of model parameters (Bearse & Bozdogan, 2000). Since a smaller value of ICOMPIFIM indicates a better model, the GA implementation in this study intends to minimize the fitness function. Models with lower ICOMPIFIM values have higher fitness scores, hence a better model-data fit. It should be noted that other information criteria such as AIC, CAIC, SBC could also be used as the fitness function for GA. Lanning (2008), Y. Liu (2007), Z. Liu and Bozdogan (2008), and Zhang (2007) have successfully combined the use of GA with information model selection criteria, ICOMP in particular.

4.3 GA Parameters

The GA initializes N randomly-selected binary strings or models to begin with. Here, N is termed population size, or the number of candidate models in the initial and subsequent pools of models. The choice of N is typically determined experimentally. In addition to population size, there are several more parameters that need to be specified before the algorithm is implemented. These parameters include number of generations, probability of crossover and probability of

mutation.

1. Number of generations: The number of times the evolutionary model creation process is repeated.

2. Probability of crossover p_c: The probability controlling the pair of chromosomes chosen for crossover.

3. Probability of mutation p_m: The probability that a randomly selected locus alternates between 0 and 1.

Besides, the algorithm allows the best model from the current GA generation to be included in the next one, which is called *elitism rule*. In this study, this rule is always applied.

4.4 The Iteration Process

Given these parameters and the initial population of models, the algorithm iteratively explores the model space through an evolutionary process. This process continues until the predefined number of generations is exhausted. Moving from one generation to another, new models are created through the operations of natural selection, crossover and mutation (Bearse & Bozdogan, 2000).

For the purpose of constructing a mechanism to iteratively improve model

selection results through an evolutionary process, this study uses the rank ordering of $(-1) * ICOMPIFIM$. The candidate models are sorted in an ascending order by the rank of $(-1) * ICOMPIFIM$. The model with the largest/worst ICOMPIFIM value has the smallest ranking 1 while that with the smallest/best ICOMPIFIM value has the highest ranking. Then a weighted roulette wheel with N bins is constructed with each bin corresponding to each subset MVR model. And the bin width for a model with *Rank* i is $\frac{i}{N(N+1)/2}$. Then N uniform distribution random numbers are generated from $(0, 1)$. Therefore, with each of the N bins or N models, there is a uniform random number associated. Each time the random number falls in its bin, the corresponding model is included in the mating pool. Since better models have wider bins, it is expected that they will be better represented in the mating pool due to the improved chance of the random number falling into the bin. By applying such a mechanism, the natural selection role of the GA is achieved (Bearse & Bozdogan, 2000).

To determine the subset MVR models included in the next GA generation, a crossover operation is applied to the mating pool which recombines subset MVR models of the current generation, or the parents, into new subset MVR models for the next generation, or the offsprings. To perform crossover, the subset MVR

models in the mating pool are randomly paired. Each locus in the binary coding scheme is swapped with the corresponding locus of its mate with crossover probability p_c. The overall probability of at least one locus crossing over in a given mating pair is given by Equation 27.

$$p_c^* = 1 - (1 - p_c)^{string\ length}, \tag{27}$$

where *string length* equals the length of the binary coding representing an MVR subset model, or the number of response variables times the number of predictors.

When $p_c = 0$, $p_c^* = 0$, which indicates the subset models in the next population are identical to those in the current one. Whenever $p_c \in (0, 1)$, the offspring models are expected to differ from the mating pool.

The resulting N offspring models are then subjected to mutation. Mutation is a means of creating new combinations of variables not available in the current area of the model space, thereby allowing the GA to create models not attainable through the crossover operation alone. Mutation is controlled by a user-defined mutation probability p_m at which a locus on the binary coding alternates between 0 and 1. Therefore, for each offspring model, the mutation technique allows a predictor variable to be added or removed randomly.

When subset models are generated by GA, they are evaluated by the fitness function under the MVR modeling situation to find a champion subset. And this champion subset is then subjected to a two-step estimation scheme to obtain the estimated regression coefficient matrix in $\mathbf{F} = \mathbf{XB} + \mathbf{E}$. Both the two-step scheme and the mathematical formulas of the GA fitness functions are covered in the section that follows.

5 Multivariate Regression

5.1 Overview of Multivariate Regression

Multivariate regression is a statistical technique that uses a set of independent variables to predict two or more response variables simultaneously. This is often seen in many areas of application including econometrics, behavioral sciences, social sciences, etc. In this study, factor scores $\widehat{\mathbf{F}}$ estimated from MLE common factor analysis are to be regressed on the original data set $\mathbf{X}$ to find out how the factor scores depend upon the original set of variables. Stated differently, an optimal or near optimal subset of original variables is to be found that could be used to predict the factor scores by estimating the regression coefficient matrix $\mathbf{B}$ in $\mathbf{F} = \mathbf{XB} + \mathbf{E}$.

To that end, information model selection criteria are employed which evaluate the fit of each MVR subset generated by the GA. The subset that minimizes the information criterion of choice, or the fitness function for the GA, is selected as the champion model. Covered in this section are the theoretical background of MVR model parameter estimation given a subset of predictors and how it

relates to the mathematical formulas of the two information criteria: $AIC\,(MVR)$ and $ICOMPIFIM\,(MVR)$. As has been mentioned in the previous section, $ICOMPIFIM\,(MVR)$ is the information criterion of choice, or it serves as the GA fitness function in the study.

5.2 Gaussian MVR Model

5.2.1 Mathematical Forms & Model Assumptions

The Gaussian MVR model under discussion is written as

$$\mathbf{F} = \mathbf{XB} + \mathbf{E}, \tag{28}$$

where

$$\underset{(n\times m)}{\mathbf{F}} = (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n)', \tag{29}$$

$$\underset{(n\times p)}{\mathbf{X}} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)', \tag{30}$$

$$\underset{(p\times m)}{\mathbf{B}} = \begin{bmatrix} \beta_{11} & ... & \beta_{1m} \\ ... & ... & ... \\ \beta_{p1} & ... & \beta_{pm} \end{bmatrix}, \tag{31}$$

$$\underset{(n\times m)}{\mathbf{E}} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, ..., \boldsymbol{\varepsilon}_n)', \tag{32}$$

$$\boldsymbol{\varepsilon}_t \tilde{} i.i.d.\ N_m\,(\mathbf{0}, \boldsymbol{\Sigma}), \tag{33}$$

$$t = 1, 2, ..., n.$$

In Equation 31, $\underset{(p\times m)}{\mathbf{B}} = \left[\beta_{ij}\right]$, $i = 1, 2, ..., p$, $j = 1, 2, ..., m$, is the matrix of regression coefficients. The matrix element β_{ij} denotes the partial effect of the ith predictor on the jth response variable. When no zero restrictions are imposed on the elements of $\mathbf{B}$, Equation 28 represents the saturated MVR model. That is to say, each of the m response variables is being predicted by all p predictors.

5.2.2 Imposing Zero Restrictions on the Coefficient Matrix

When zero restrictions are imposed on the elements of $\mathbf{B}$, subset MVR models are obtained. In order to identify from all GA-generated MVR subsets the optimal one that achieves satisfactory predictive power and a reasonable level of parsimony, information criteria are to be scored that map the fit of an MVR model to a scalar value. As a consequence, the study proceeds to derive mathematical formulas of the information criteria for the MVR model.

To that end, the model is rewritten in the vectorized notation:

$$\begin{aligned} vec\left(\mathbf{F}\right) &= vec\left(\mathbf{XB}\right) + vec\left(\mathbf{E}\right) \\ &= \left(\mathbf{I}_m \otimes \mathbf{X}\right) vec\left(\mathbf{B}\right) + vec\left(\mathbf{E}\right), \end{aligned} \tag{34}$$

or

$$\underset{(nm\times 1)}{\mathbf{f}} = \underset{(nm\times mp)}{\mathbf{X}_{\sup}} \underset{(mp\times 1)}{\boldsymbol{\beta}} + \underset{(nm\times 1)}{\mathbf{e}}, \tag{35}$$

$$\mathbf{e}\tilde{} N_{nm}\left(\mathbf{0}, \boldsymbol{\Omega}\right), \tag{36}$$

where

$$\mathbf{f} = vec\left(\mathbf{F}\right), \tag{37}$$

$$\mathbf{X}_{\sup} = \left(\mathbf{I}_m \otimes \mathbf{X}\right), \tag{38}$$

$$\boldsymbol{\beta} = vec\left(\mathbf{B}\right), \tag{39}$$

$$\mathbf{e} = vec\left(\mathbf{E}\right), \tag{40}$$

$$\underset{(nm\times nm)}{Cov\left(\mathbf{e}\right)} = \boldsymbol{\Omega} = \left(\boldsymbol{\Sigma} \otimes \mathbf{I}_n\right), \tag{41}$$

where $\otimes$ denotes the *Kronecker product*,

$$\underset{(m\times m)}{\boldsymbol{\Sigma}} = Cov\left(\boldsymbol{\varepsilon}_t\right), \tag{42}$$

$$t = 1,\ 2,\ ...,\ n.$$

Let

$$\underset{(mp\times 1)}{\boldsymbol{\beta}} = \underset{(mp\times w)}{\mathbf{R}} \underset{(w\times 1)}{\boldsymbol{\gamma}}, \tag{43}$$

where $\mathbf{R}$ is a matrix consisting of only 0's and 1's,

$$w \leq mp. \tag{44}$$

In Equation 44, w is the number of nonzero or unrestricted elements in the $(mp \times 1)$ vector $\boldsymbol{\beta}$. So, it holds that $w = (mp - Number\ of\ zero\ restrictions\ on\ \boldsymbol{\beta})$. The $(w \times 1)$ vector $\boldsymbol{\gamma}$ contains all the unrestricted elements in $\boldsymbol{\beta}$.

If the ith element of $\boldsymbol{\beta}$ is restricted to 0, the ith row of $\mathbf{R}$ consists of all 0's with no exceptions whatsoever. If the ith element of $\boldsymbol{\beta}$ is unrestricted, the ith row of $\mathbf{R}$ consists of all 0's with the exception of the jth column, which has a one in it, where j equals the total number of unrestricted $\boldsymbol{\beta}$ elements at and prior to the current row. Followed next is a demonstration of the determination of j if the ith element of $\boldsymbol{\beta}$ is unrestricted.

If the following indicator variable is used:

$$u_i = \begin{cases} 1 & \text{if the } ith \text{ row of } \boldsymbol{\beta} \text{ is unrestricted} \\ 0 & \text{if the } ith \text{ row of } \boldsymbol{\beta} \text{ is restricted} \end{cases}, \tag{45}$$

then $j = \sum_{i \leq k} u_i$ if the kth element of $\boldsymbol{\beta}$ is unrestricted.

In order to facilitate a better understanding of how the choice of $\mathbf{R}$ and $\boldsymbol{\gamma}$ relates to $\boldsymbol{\beta}$ and MVR subsets, an example is presented below with $m = 2$, $p = 3$. Then

$$\underset{(p \times m)}{\mathbf{B}} = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \\ \beta_{31} & \beta_{32} \end{bmatrix}. \tag{46}$$

$$\begin{bmatrix} f_{t1} \\ f_{t2} \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{21} & \beta_{31} \\ \beta_{12} & \beta_{22} & \beta_{32} \end{bmatrix} \begin{bmatrix} x_{t1} \\ x_{t2} \\ x_{t3} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t1} \\ \varepsilon_{t2} \end{bmatrix}, \tag{47}$$

where $t = 1, 2, ..., n$.

Based on Equation 39, the following result is obtained:

$$\boldsymbol{\beta} = vec\,(\mathbf{B}) = [\beta_{11}\ \beta_{21}\ \beta_{31}\ \beta_{12}\ \beta_{22}\ \beta_{32}]'. \tag{48}$$

Suppose the restrictions imposed on the elements of $\mathbf{B}$ are

$$\beta_{21} = 0, \tag{49}$$

$$\beta_{31} = 0, \tag{50}$$

$$\beta_{22} = 0. \tag{51}$$

Then

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{11} \\ 0 \\ 0 \\ \beta_{12} \\ 0 \\ \beta_{32} \end{bmatrix}. \tag{52}$$

By doing so, the predictors x_2 and x_3 are excluded from the equation for predicting f_1, and the predictor x_2 is removed from the equation for predicting f_2.

In order to obtain Equation 52 using Equation 43, the construction of $\mathbf{R}$ and $\boldsymbol{\gamma}$ should be

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{53}$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{32} \end{bmatrix}. \tag{54}$$

Then

$$\boldsymbol{\beta} = \mathbf{R}\boldsymbol{\gamma} = \begin{bmatrix} \beta_{11} \\ 0 \\ 0 \\ \beta_{12} \\ 0 \\ \beta_{32} \end{bmatrix}, \text{ which is the desired outcome.} \tag{55}$$

By the appropriate choice of the $\mathbf{R}$ matrix, zero restrictions can be imposed on the elements of the $\mathbf{B}$ matrix, which in turn yields different MVR subsets. Then, $\boldsymbol{\beta}$ in Equation 35 is replaced with $\mathbf{R}$ and $\boldsymbol{\gamma}$ based on Equation 55 so that the MVR model is transformed to a multiple regression model where the number of response variables is 1 and the regression coefficient matrix is $\boldsymbol{\gamma}$. With $\mathbf{R}$ and an estimate of $\boldsymbol{\gamma}$, an estimate of $\boldsymbol{\beta}$ can be obtained based on Equation 55 and it is then restructured to derive an estimate of $\mathbf{B}$ according to Equation 46 and Equation 48.

Considering both Equation 35 and Equation 43, the following results can be obtained.

$$\begin{aligned} \underset{(nm\times 1)}{\mathbf{f}} &= \underset{(nm\times mp)}{\mathbf{X}_{\text{sup}}}\underset{(mp\times 1)}{\boldsymbol{\beta}} + \underset{(nm\times 1)}{\mathbf{e}} \\ &= \underset{(nm\times mp)}{\mathbf{X}_{\text{sup}}}\underset{(mp\times w)}{\mathbf{R}}\underset{(w\times 1)}{\boldsymbol{\gamma}} + \underset{(nm\times 1)}{\mathbf{e}} \\ &= \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}}\underset{(w\times 1)}{\boldsymbol{\gamma}} + \underset{(nm\times 1)}{\mathbf{e}}, \end{aligned} \tag{56}$$

or

$$\underset{(nm\times 1)}{\mathbf{f}} = \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}} \underset{(w\times 1)}{\boldsymbol{\gamma}} + \underset{(nm\times 1)}{\mathbf{e}}, \tag{57}$$

where $\mathbf{X}^{*}_{\text{sup}}$ contains the predictors for the subset model. This is because $\mathbf{X}^{*}_{\text{sup}}$ has w columns and $\mathbf{X}_{\text{sup}}$ mp columns. Based on Equation 44, $w \leq mp$.

With the help of the $\mathbf{R}$ matrix, the subset MVR model in Equation 28, with zero restrictions imposed on the elements of $\mathbf{B}$, has been transformed to a multiple regression model in Equation 57 where the dimension of the observed response variable matrix $\mathbf{f}$ is $(nm \times 1)$ and that of the observed independent variable matrix $\mathbf{X}^{*}_{\text{sup}}$ is $(nm \times w)$.

5.2.3 A Two-Step Approach to MVR Parameter Estimation

For a multiple regression model in Equation 57 which is originally derived from a subset MVR model, a two-step estimation scheme is employed to obtain *feasible generalized least squares* $(FGLS)$ estimates for $\boldsymbol{\gamma}$, its covariance matrix $Cov\,(\boldsymbol{\gamma})$, and the covariance matrix of $\mathbf{e}$ residuals $\boldsymbol{\Omega} \equiv Cov\,(\mathbf{e}) \equiv \boldsymbol{\Sigma} \otimes \mathbf{I}_n$.

Step 1. Obtain a consistent estimator of $\boldsymbol{\Omega} \equiv Cov\,(\mathbf{e})$

a. Construct the *least squares* (LS) *estimator*

$$\underset{(w\times 1)}{\widehat{\boldsymbol{\gamma}}} = \left(\underset{(w\times nm)}{\mathbf{X}^{*\prime}_{\text{sup}}} \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}} \right)^{-1} \underset{(w\times nm)}{\mathbf{X}^{*\prime}_{\text{sup}}} \underset{(nm\times 1)}{\mathbf{f}}. \tag{58}$$

b. Construct a consistent estimator of $\boldsymbol{\varepsilon}$ from

$$\underset{(nm\times 1)}{\widehat{\boldsymbol{\varepsilon}}} = vec\left(\underset{(n\times m)}{\widehat{\mathbf{E}}} \right) = \underset{(nm\times 1)}{\mathbf{f}} - \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}} \underset{(w\times 1)}{\widehat{\boldsymbol{\gamma}}}. \tag{59}$$

Let $\widehat{\varepsilon}_i$ denote the ith element of $\widehat{\boldsymbol{\varepsilon}}$ and define

$$\underset{(n\times m)}{\widehat{\mathbf{E}}} = \begin{bmatrix} \widehat{\varepsilon}_1 & \widehat{\varepsilon}_{n+1} & \cdots & \widehat{\varepsilon}_{n(m-1)+1} \\ \widehat{\varepsilon}_2 & \widehat{\varepsilon}_{n+2} & \cdots & \widehat{\varepsilon}_{n(m-1)+2} \\ \cdots & \cdots & \cdots & \cdots \\ \widehat{\varepsilon}_n & \widehat{\varepsilon}_{2(n)} & \cdots & \widehat{\varepsilon}_{n(m)} \end{bmatrix} = \begin{bmatrix} \widehat{\varepsilon}_{1,1} & \widehat{\varepsilon}_{1,2} & \cdots & \widehat{\varepsilon}_{1,m} \\ \widehat{\varepsilon}_{2,1} & \widehat{\varepsilon}_{2,2} & \cdots & \widehat{\varepsilon}_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ \widehat{\varepsilon}_{n,1} & \widehat{\varepsilon}_{n,2} & \cdots & \widehat{\varepsilon}_{n,m} \end{bmatrix}. \tag{60}$$

c. Construct a consistent estimator of $\boldsymbol{\Sigma}$ by

$$\underset{(m\times m)}{\widehat{\boldsymbol{\Sigma}}} = \frac{1}{n} \underset{(m\times n)}{\widehat{\mathbf{E}}^{\prime}} \underset{(n\times m)}{\widehat{\mathbf{E}}}. \tag{61}$$

d. Construct a consistent estimator of $\boldsymbol{\Omega}$ from

$$\underset{(nm\times nm)}{\widehat{\boldsymbol{\Omega}}} = \underset{(m\times m)}{\widehat{\boldsymbol{\Sigma}}} \otimes \underset{(n\times n)}{\mathbf{I}_n}. \tag{62}$$

Step 2. Obtain the *FGLS estimator* of $\boldsymbol{\gamma}$ & $\boldsymbol{\Omega}$

a. Construct the *FGLS estimator* of $\boldsymbol{\gamma}$

$$\underset{(w\times 1)}{\widetilde{\boldsymbol{\gamma}}} = \left(\underset{(w\times nm)}{\mathbf{X}^{*\prime}_{\text{sup}}} \underset{(nm\times nm)}{\widehat{\boldsymbol{\Omega}}^{-1}} \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}} \right)^{-1} \underset{(w\times nm)}{\mathbf{X}^{*\prime}_{\text{sup}}} \underset{(nm\times nm)}{\widehat{\boldsymbol{\Omega}}^{-1}} \underset{(nm\times 1)}{\mathbf{f}}. \tag{63}$$

b. Construct the *FGLS residuals*

$$\underset{(nm\times 1)}{\widetilde{\boldsymbol{\varepsilon}}} = vec\left(\underset{(n\times m)}{\widetilde{\mathbf{E}}} \right) = \underset{(nm\times 1)}{\mathbf{f}} - \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}} \underset{(w\times 1)}{\widetilde{\boldsymbol{\gamma}}}. \tag{64}$$

c. Construct the *FGLS estimator* of $\boldsymbol{\Sigma}$ by reshaping the previous equation

$$\underset{(m\times m)}{\widetilde{\boldsymbol{\Sigma}}} = \frac{1}{n} \underset{(m\times n)}{\widetilde{\mathbf{E}}^{\prime}} \underset{(n\times m)}{\widetilde{\mathbf{E}}}. \tag{65}$$

d. Construct the *FGLS estimator* of $\boldsymbol{\Omega}$ from

$$\underset{(nm\times nm)}{\widetilde{\boldsymbol{\Omega}}} = \underset{(m\times m)}{\widetilde{\boldsymbol{\Sigma}}} \otimes \underset{(n\times n)}{\mathbf{I}_n}. \tag{66}$$

Under Gaussian errors, the following things should be noted.

- The *ML estimators* of $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ are the *GLS estimators*.
- The *FGLS* estimators have the same asymptotic distributions as the *GLS* estimators, and consequently, the *ML* estimators.

For a model in Equation 57, after the two-step estimation scheme, several estimates are obtained.

$$\underset{(w\times 1)}{\widetilde{\boldsymbol{\gamma}}} = \left(\underset{(w\times nm)}{\mathbf{X}^{*\prime}_{\text{sup}}} \underset{(nm\times nm)}{\widetilde{\boldsymbol{\Omega}}^{-1}} \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}} \right)^{-1} \underset{(w\times nm)}{\mathbf{X}^{*\prime}_{\text{sup}}} \underset{(nm\times nm)}{\widetilde{\boldsymbol{\Omega}}^{-1}} \underset{(nm\times 1)}{\mathbf{f}}. \tag{67}$$

A consistent estimator of the covariance matrix of $\widetilde{\boldsymbol{\gamma}}$ is given by

$$\underset{(w\times w)}{\widehat{Cov}\left(\widetilde{\boldsymbol{\gamma}}\right)} = \left(\underset{(w\times nm)}{\mathbf{X}^{*\prime}_{\text{sup}}} \underset{(nm\times nm)}{\widetilde{\boldsymbol{\Omega}}^{-1}} \underset{(nm\times w)}{\mathbf{X}^{*}_{\text{sup}}} \right)^{-1}. \tag{68}$$

Similarly a consistent estimator of the covariance matrix of $\widetilde{\boldsymbol{\varepsilon}}$ residuals is given by

$$\underset{(nm\times nm)}{\widehat{Cov}\left(\widetilde{\boldsymbol{\varepsilon}}\right)} = \underset{(nm\times nm)}{\widetilde{\boldsymbol{\Omega}}} = \underset{(m\times m)}{\widetilde{\boldsymbol{\Sigma}}} \otimes \underset{(n\times n)}{\mathbf{I}_n}. \tag{69}$$

Based on the results from the two-step estimation scheme, the log likelihood function for the subset MVR model under Gaussian errors,

$$
\begin{aligned}
\log L(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) &= -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) \\
&\quad -\frac{1}{2}\left[(\mathbf{f} - (\mathbf{I}_m \otimes \mathbf{X})\boldsymbol{\gamma})'\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n\right)(\mathbf{f} - (\mathbf{I}_m \otimes \mathbf{X})\boldsymbol{\gamma})\right] \\
&= -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) \\
&\quad -\frac{1}{2}tr\left[(\mathbf{F} - \mathbf{XB})'(\mathbf{F} - \mathbf{XB})\boldsymbol{\Sigma}^{-1}\right] \\
&= -\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{nm}{2},
\end{aligned} \tag{70}
$$

can now be maximized.

5.2.4 Information Criteria for MVR Subset Selection

Therefore, according to the results from the two-step scheme and the maximized log likelihood function for a Gaussian MVR model, the analytical forms of information model selection criteria $AIC(MVR)$ and $ICOMPIFIM(MVR)$ can be derived. The open-form formulas are given below. It should be noted that the required inputs for the two formulas are available as part of the standard output of most regression packages.

$$
AIC(MVR) = nm\log(2\pi) + n\log\left(\left|\widetilde{\boldsymbol{\Sigma}}\right|\right) + nm + 2\left[w + \frac{m(m+1)}{2}\right], \tag{71}
$$

$$ICOMPIFIM\left(MVR\right)=nm\log\left(2\pi\right)+n\log\left(\left|\widetilde{\boldsymbol{\Sigma}}\right|\right)+nm \tag{72}$$
$$+s\log\left\{\frac{tr\left[\widehat{Cov}\left(\widetilde{\boldsymbol{\gamma}}\right)\right]+\frac{1}{n}\left[\frac{1}{2}tr\left(\widetilde{\boldsymbol{\Sigma}}^{2}\right)+\frac{1}{2}tr^{2}\left(\widetilde{\boldsymbol{\Sigma}}\right)+\sum_{j=1}^{m}\left(\left(\widetilde{\sigma}_{jj}^{2}\right)^{2}\right)\right]}{s}\right\}$$
$$-\log\left|\widehat{Cov}\left(\widetilde{\boldsymbol{\gamma}}\right)\right|-m\log\left(2\right)+\frac{m\left(m+1\right)}{2}\log\left(n\right)-\left(m+1\right)\log\left|\widetilde{\boldsymbol{\Sigma}}\right|,$$

where

$$s=\dim\left(\widehat{F}^{-1}\right)=rank\left(\widehat{F}^{-1}\right), \tag{73}$$

$$\widehat{F}^{-1}\equiv\widehat{F}^{-1}\left(\widetilde{\boldsymbol{\gamma}},\widetilde{\boldsymbol{\Sigma}}\right)=\begin{bmatrix}\underset{(w\times w)}{\widehat{Cov}\left(\widetilde{\boldsymbol{\gamma}}\right)} & \mathbf{0}\\ \mathbf{0}' & \frac{2}{n}\mathbf{D}_{m}^{+}\left(\widetilde{\boldsymbol{\Sigma}}\otimes\widetilde{\boldsymbol{\Sigma}}\right)\mathbf{D}_{m}^{+\prime}\\ & \left(\frac{m\left(m+1\right)}{2}\times\frac{m\left(m+1\right)}{2}\right)\end{bmatrix}, \tag{74}$$

$\mathbf{D}_{m}^{+}$ is the $Moore-Penrose\ inverse$ of the duplication matrix $\mathbf{D}_{m}$,

$$\mathbf{D}_{m}^{+}=\left(\mathbf{D}_{m}^{'}\mathbf{D}_{m}\right)^{-1}\mathbf{D}_{m}^{'}, \tag{75}$$

and $\widetilde{\sigma}_{jj}^{2}$, $j=1,\ 2,\ ...,\ m$, is the jth diagonal element of $\underset{(m\times m)}{\widetilde{\boldsymbol{\Sigma}}}$.

With $\widetilde{\boldsymbol{\gamma}}$ and $\mathbf{R}$, $\widetilde{\boldsymbol{\beta}}$ can be obtained according to Equation 55, so $\widetilde{\mathbf{B}}$ can be derived by restructuring $\widetilde{\boldsymbol{\beta}}$ based on Equation 46 and Equation 48. $\widetilde{\mathbf{B}}$ will then be sparsed to obtain $\widetilde{\mathbf{B}}_{Sparsed}$, which means any value in $\widetilde{\mathbf{B}}$ that falls between

the absolute value of the corresponding column mean and its negative value is considered to be a zero. The pattern of zeroes in $\widetilde{\mathbf{B}}_{Sparsed}$ is used to *zero out* corresponding elements in the factor loading matrix to obtain the best factor pattern structure being sought, and the structure is then used as an initial pattern for a confirmatory factor analysis.

Now with the information already presented previously, the following goals can be achieved.

1. Choosing the number of factors in maximum likelihood factor analysis using the information criterion of choice, obtaining the factor solution and estimating factor scores.

2. Choosing the optimal MVR subset of original variables to predict factor scores using the GA with an ICOMP-type criterion as the fitness function.

3. Estimating MVR parameters given the GA-selected optimal MVR subset to establish the factor pattern structure that is being sought.

4. Sparsing $\widetilde{\mathbf{B}}$ to obtain $\widetilde{\mathbf{B}}_{Sparsed}$ which is then used to zero out the factor loading matrix for the purpose of deriving the best factor pattern structure.

In each of the following five examples, $\widetilde{\mathbf{B}}$ will be obtained. In the first two examples, $\widetilde{\mathbf{B}}$ will be further sparsed to determine the best factor pattern structure.

And several relevant correlation coefficient matrices will be estimated to check some factor model assumptions presented previously. It should be noted that the sparsing procedure could also be easily applied to the other three examples upon request.

6 Numerical Examples

In this section, several data sets are analyzed. The outcomes from the analyses demonstrate the application of the new approach to factor analysis using information criteria and Genetic Algorithm. In order to save space, some abbreviations for ICOMP-type criteria are used in the study. $ICOMPIFIM$ is abbreviated as $ICOMP1$, $ICOMPIFIM_{PEU\_Mis}$ as $ICOMP2$, and $ICOMPIFIM_{PEU\_Mis\_LN}$ as $ICOMP3$. GA is used in analyzing each example. The GA population size is 20 and the number of generations is 30. The probability of crossover is 0.50 and the probability of mutation is 0.01. ICOMPIFIM, as is defined in Equation 72, serves as the fitness function and the elitism rule is always applied.

6.1 Example 1. *A Simulation Study*

This data set consists of 100 *i.i.d.* observations and is simulated from a 12-dimensional multivariate normal distribution (Bozdogan & Ramirez, 1987). The multivariate normal distribution has a zero mean vector and a covariance matrix $\mathbf{\Sigma}_{Sim} = \mathbf{\Lambda}_{Sim}\mathbf{\Lambda}'_{Sim} + \mathbf{\Psi}_{Sim}$. Here, $\mathbf{\Lambda}_{Sim}$ and $\mathbf{\Psi}_{Sim}$ are set up in the following way:

$$\underset{(12\times 3)}{\mathbf{\Lambda}_{Sim}} = \begin{bmatrix} .9 & 0 & 0 \\ .9 & 0 & 0 \\ .9 & 0 & 0 \\ .9 & 0 & 0 \\ 0 & .8 & 0 \\ 0 & .8 & 0 \\ 0 & .8 & 0 \\ 0 & .8 & 0 \\ 0 & 0 & .7 \\ 0 & 0 & .7 \\ 0 & 0 & .7 \\ 0 & 0 & .7 \end{bmatrix}, \tag{76}$$

and

$$\underset{(12\times 12)}{\mathbf{\Psi}_{Sim}} = Diag\left(.19, .19, .19, .19, .36, .36, .36, .36, .51, .51, .51, .51\right). \tag{77}$$

When the data is simulated, maximum likelihood factor analysis is run using $MATLAB^{\text{TM}}$'s $factoran$ function. When there are 12 variables, this function can fit up to 7 factors. And as is described above, for each fitted factor model, six information criteria are scored for the purpose of evaluating the model-data fit. Table 1 has in it the information criterion scores for all fitted factor models.

$ICOMP1$ or $ICOMPIFIM$ is minimized at 2785.1 when $m = 3$. So the best approximating factor model is selected as the one with 3 factors. As a result, Genetic Algorithm is run for the 3-factor model.

Using previously specified parameters, GA is run three times and each run of

the GA leads to a minimized fitness function or $ICOMPIFIM$. A total of three minimum $ICOMPIFIM$ values from three runs of GA are obtained. Then the champion MVR subset is selected as the one that corresponds to the smallest of these three minimum values.

Presented in Figure 1, Figure 2, and Figure 3 are the progress graphs for the three runs of GA. In each graph, average and minimum fitness function values are plotted against GA generation index, respectively, with the one on top corresponding to the average ICOMPIFIM value for that GA generation and the other one the minimum ICOMPIFIM value.

As can be seen from Figure 1, Figure 2, and Figure 3, the information criterion ICOMPIFIM decreases substantially as the GA moves from one generation to another. Since a smaller value of ICOMPIFIM indicates a better model, the graphs show that GA is capable of finding better models through an iterative process. Table 2 shows the selected optimal MVR subset model from the GA. For this subset, ICOMPIFIM is minimized at -3042.8. Note that the columns represent factors and the rows original variables.

When all GA generations are finished, the subset MVR model for the model $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ is obtained where $\mathbf{F}$ is an n by m matrix of estimated factor scores

from maximum likelihood factor analysis and $\mathbf{X}$ is an n by p matrix of original data. From the application of GA, already obtained is to what subset of original variables each response variable or extracted factor is related.

Next comes the use of *feasible generalized least squares* ($FGLS$) method to estimate the regression coefficient matrix $\mathbf{B}$ for the purpose of identifying the complex relationship between extracted factors and original variables. Stated differently, given $\mathbf{F}$, $\mathbf{X}$ and the inclusion/exclusion information of a predictor in $\mathbf{B}$, the regression weight matrix $\mathbf{B}$ is to be estimated by regressing $\mathbf{F}$ on $\mathbf{X}$ using the $FGLS$ method. The estimated matrix $\widetilde{\mathbf{B}}$ is a p by m matrix. In this problem, $p = 12$ and $m = 3$. So $\widetilde{\mathbf{B}}$ should be 12 by 3.

$$\underset{(12\times 3)}{\widetilde{\mathbf{B}}} = \begin{bmatrix} 0.3802 & \mathbf{0.0861} & \mathbf{-0.0191} \\ 0.2073 & \mathbf{0.1005} & \mathbf{-0.0065} \\ 0.1890 & \mathbf{0.0882} & \mathbf{0.0196} \\ 0.2644 & \mathbf{0.0555} & \mathbf{-0.0618} \\ \mathbf{-0.0584} & 0.2863 & \mathbf{-0.0128} \\ \mathbf{-0.0393} & 0.2018 & \mathbf{-0.0005} \\ \mathbf{-0.0410} & 0.2730 & \mathbf{-0.0147} \\ \mathbf{-0.0222} & 0.4181 & -0.1413 \\ \mathbf{0.0104} & \mathbf{0.0369} & 0.2767 \\ \mathbf{0.0034} & \mathbf{0.0662} & 0.4049 \\ 0 & \mathbf{-0.0059} & 0.3293 \\ \mathbf{-0.0117} & 0 & 0.4792 \end{bmatrix}. \tag{78}$$

To find out about how well $\widetilde{\mathbf{B}}$ relates $\mathbf{F}$ to $\mathbf{X}$, it is natural to observe the

residuals: $\widehat{\mathbf{F}}-\widetilde{\mathbf{F}} = \widehat{\mathbf{F}}-\widetilde{\mathbf{B}}\mathbf{X}$. Figure 4 has in it a histogram of the vectorized residuals which are pretty small, ranging from -0.1 to 0.2. In addition, the residuals cluster around 0 and are approximately normally distributed. This indicates that the multivariate regression model that has been built provides a good fit.

Next, $\widetilde{\mathbf{B}}$ is to be sparsed to obtain $\widetilde{\mathbf{B}}_{Sparsed}$. Any element in Equation 78 that falls between the absolute value of the corresponding column mean and its negative value is zeroed out. The elements in Equation 78 that meet the said criterion are in bold. Therefore,

$$\underset{(12\times 3)}{\widetilde{\mathbf{B}}_{Sparsed}} = \begin{bmatrix} 0.3802 & 0 & 0 \\ 0.2073 & 0 & 0 \\ 0.1890 & 0 & 0 \\ 0.2644 & 0 & 0 \\ 0 & 0.2863 & 0 \\ 0 & 0.2018 & 0 \\ 0 & 0.2730 & 0 \\ 0 & 0.4181 & -0.1413 \\ 0 & 0 & 0.2767 \\ 0 & 0 & 0.4049 \\ 0 & 0 & 0.3293 \\ 0 & 0. & 0.4792 \end{bmatrix}. \tag{79}$$

Based on the pattern of zeroes in $\widetilde{\mathbf{B}}_{Sparsed}$, the optimal factor pattern structure can be determined by zeroing out corresponding elements in $\widehat{\mathbf{\Lambda}}_z$. Therefore, the following is obtained:

$$\widehat{\mathbf{\Lambda}}_{zSparsed} = \begin{bmatrix} 0.9380 & 0 & 0 \\ 0.8740 & 0 & 0 \\ 0.8700 & 0 & 0 \\ 0.9148 & 0 & 0 \\ 0 & 0.7865 & 0 \\ 0 & 0.7362 & 0 \\ 0 & 0.7876 & 0 \\ 0 & 0.8163 & 0 \\ 0 & 0 & 0.6260 \\ 0 & 0 & 0.7065 \\ 0 & 0 & 0.6278 \\ 0 & 0 & 0.7392 \end{bmatrix}. \tag{80}$$

As can be seen from Equation 80, the true underlying 3-factor structure has been identified successfully. The result from the simulated data is supportive of the use of the new factor pattern search algorithm in analyzing data sets coming from the real world.

Using previously estimated factor scores $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{\Lambda}}_{zSparsed}$, residuals $\boldsymbol{\varepsilon}$ in Equation 2 can be estimated. In matrix form, those estimated residuals are noted as $\underset{(n\times p)}{\widehat{\mathbf{Er}}}$. Then three sets of correlation coefficients are estimated, namely correlations between estimated residuals and estimated factor scores $corr\left(\widehat{\mathbf{Er}}, \widehat{\mathbf{F}}\right)$, correlations between original variables and estimated factor scores $corr\left(\mathbf{X}, \widehat{\mathbf{F}}\right)$, and interfactor correlations $corr\left(\widehat{\mathbf{F}}, \widehat{\mathbf{F}}\right)$.

$$corr\left(\widehat{\mathbf{Er}}, \widehat{\mathbf{F}}\right) = \begin{bmatrix} 0.0234 & 0.2864 & 0.0009 \\ 0.0060 & 0.3896 & 0.0054 \\ 0.0139 & 0.3873 & 0.0490 \\ 0.0021 & 0.3173 & -0.1212 \\ -0.4965 & 0.0040 & -0.0480 \\ -0.3857 & 0.0034 & -0.0167 \\ -0.3797 & 0.0061 & -0.0476 \\ -0.5510 & 0.0406 & -0.0040 \\ 0.0322 & 0.1972 & -0.0125 \\ -0.2113 & 0.3015 & -0.0139 \\ 0.3020 & 0.0875 & 0.0124 \\ -0.0094 & 0.0895 & 0.0012 \end{bmatrix}. \quad (81)$$

As can be seen from Equation 81, most of the correlation coefficients are very close to 0, indicating that the correlations between the residuals and the extracted factors are almost nonexistent. Therefore, the assumption outlined in Equation 9 is satisfied.

$$corr\left(\mathbf{X}, \widehat{\mathbf{F}}\right) = \begin{bmatrix} 0.9608 & 0.0663 & 0.0000 \\ 0.8921 & 0.1639 & 0.0022 \\ 0.8916 & 0.1633 & 0.0220 \\ 0.9316 & 0.1026 & -0.0443 \\ -0.2899 & 0.8278 & -0.0271 \\ -0.2544 & 0.7749 & -0.0107 \\ -0.2231 & 0.8301 & -0.0268 \\ -0.2681 & 0.8757 & -0.1312 \\ 0.0229 & 0.1412 & 0.6976 \\ -0.1301 & 0.1851 & 0.7889 \\ 0.2103 & 0.0608 & 0.7173 \\ -0.0054 & 0.0490 & 0.8349 \end{bmatrix}. \quad (82)$$

As can be seen from Equation 82, the pattern of positive and negative values

in this matrix is almost the same as that in the factor loading matrix with the exception of only three matrix elements, which is supportive of the effectiveness of the techniques presented in the study.

$$corr\left(\widehat{\mathbf{F}},\widehat{\mathbf{F}}\right) = \begin{bmatrix} 1.0000 & -0.0137 & -0.0002 \\ -0.0137 & 1.0000 & -0.0002 \\ -0.0002 & -0.0002 & 1.0000 \end{bmatrix}. \tag{83}$$

As can be seen from Equation 83, all correlation coefficients are very close to zero, indicating that the extracted factors are statistically independent and orthogonal of each other.

Figure 5 presents the optimal factor pattern structure that has just been identified. All factor loadings in the figure are available in Equation 80.

6.2 Example 2. *Kendall Job Applicant Data*

This data set comes from scores of 48 job applicants for a certain job in UK. In this data set, each of the 48 applicants is measured on 15 variables. So the final data used for the regression model $\mathbf{F} = \mathbf{X}\ \mathbf{B} + \mathbf{E}$ has a total of 15 variables, or $p = 15$. Background information of the 15 variables is presented in Table 3.

Next, maximum likelihood factor analysis is run using $MATLAB^{\text{TM}}$'s $factoran$ function. When there are 15 variables, this function can fit up to 10 factors. And

as is described above, for each fitted factor model, six information criteria are scored for the purpose of evaluating the model-data fit. Table 4 has in it the information criterion scores for all fitted factor models.

$ICOMP1$ or $ICOMPIFIM$ is minimized at 1689.1 when $m = 5$. So the best approximating factor model is selected as the one with 5 factors. As a result, Genetic Algorithm is run for the 5-factor model.

The GA parameters used for this data set are identical to those for the previous one. GA is run three times and each run of the GA leads to a minimized fitness function or $ICOMPIFIM$. A total of three minimum $ICOMPIFIM$ values from three runs of GA are obtained. Then the champion MVR subset is selected as the one that corresponds to the smallest of these three minimum values.

Presented in Figure 6, Figure 7, and Figure 8 are the progress graphs for the three runs of GA. In each graph, average and minimum fitness function values are plotted against GA generation index, respectively, with the one on top corresponding to the average ICOMPIFIM value for that GA generation and the other one the minimum ICOMPIFIM value.

As can be seen from Figure 6, Figure 7, and Figure 8, the information criterion ICOMPIFIM decreases substantially as the GA moves from one generation to

another. Since a smaller value of ICOMPIFIM indicates a better model, the graphs show that GA is capable of finding better models through an iterative process. Table 5 shows the selected optimal MVR subset model from the GA. For this subset, ICOMPIFIM is minimized at -1858.6. Note that the columns represent factors and the rows original variables.

When all GA generations are finished, the subset MVR model for the model $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ is obtained where $\mathbf{F}$ is an n by m matrix of estimated factor scores from maximum likelihood factor analysis and $\mathbf{X}$ is an n by p matrix of original data. From the application of GA, already obtained is to what subset of original variables each response variable or extracted factor is related.

Next comes the use of *feasible generalized least squares* ($FGLS$) method to estimate the regression coefficient matrix $\mathbf{B}$ for the purpose of identifying the complex relationship between extracted factors and original variables. Stated differently, given $\mathbf{F}$, $\mathbf{X}$ and the inclusion/exclusion information of a predictor in $\mathbf{B}$, the regression weight matrix $\mathbf{B}$ is to be estimated by regressing $\mathbf{F}$ on $\mathbf{X}$ using the $FGLS$ method. The estimated matrix $\widetilde{\mathbf{B}}$ is a p by m matrix. In this problem, $p = 15$ and $m = 5$. So $\widetilde{\mathbf{B}}$ should be 15 by 5.

$$\underset{(15\times5)}{\widetilde{\mathbf{B}}} = \begin{bmatrix} -0.0568 & 0 & \mathbf{0.0017} & -0.1079 & 0 \\ -0.0824 & 0 & 0.0048 & 0 & 0.0299 \\ -0.1435 & -0.0113 & 0.0124 & -0.0670 & 0.0512 \\ -0.0351 & -0.0601 & 0 & 0.1273 & 0 \\ -0.0557 & 0 & 0.0153 & 0.0905 & -0.2678 \\ 0.1410 & -0.4103 & 0.2892 & -0.3205 & 0.2807 \\ 0 & 0 & \mathbf{-0.0012} & 0.1862 & 0 \\ 0.0552 & 0.0724 & 0.0041 & 0 & -0.1488 \\ 0 & 0.0526 & \mathbf{0.0004} & -0.0663 & 0 \\ -0.0388 & 0.0327 & 0.0040 & -0.0330 & -0.0560 \\ 0 & 0.1549 & 0.0084 & 0 & -0.2383 \\ -0.0277 & 0 & 0.0127 & 0.0405 & 0 \\ 0.1694 & 0.2734 & \mathbf{-0.0014} & 0.3250 & 0.2122 \\ 0.1401 & -0.1367 & -0.3990 & -0.1685 & 0 \\ 0.0106 & 0.1013 & \mathbf{-0.0014} & -0.0728 & 0.1229 \end{bmatrix}. \quad (84)$$

To find out about how well $\widetilde{\mathbf{B}}$ relates $\mathbf{F}$ to $\mathbf{X}$, it is natural to observe the residuals: $\widehat{\mathbf{F}}-\widetilde{\mathbf{F}} = \widehat{\mathbf{F}}-\widetilde{\mathbf{B}}\mathbf{X}$. Figure 9 has in it a histogram of the vectorized residuals which are pretty small, ranging from -0.6 to 0.8. In addition, the residuals cluster around 0 and are approximately normally distributed. This indicates that the multivariate regression model that has been built provides a good fit.

Next, $\widetilde{\mathbf{B}}$ is to be sparsed to obtain $\widetilde{\mathbf{B}}_{Sparsed}$. Any element in Equation 84 that falls between the absolute value of the corresponding column mean and its negative value is zeroed out. The elements in Equation 84 that meet the said criterion are in bold. Therefore,

$$\underset{(15\times5)}{\widetilde{\mathbf{B}}_{Sparsed}} = \begin{bmatrix} -0.0568 & 0 & 0 & -0.1079 & 0 \\ -0.0824 & 0 & 0.0048 & 0 & 0.0299 \\ -0.1435 & -0.0113 & 0.0124 & -0.0670 & 0.0512 \\ -0.0351 & -0.0601 & 0 & 0.1273 & 0 \\ -0.0557 & 0 & 0.0153 & 0.0905 & -0.2678 \\ 0.1410 & -0.4103 & 0.2892 & -0.3205 & 0.2807 \\ 0 & 0 & 0 & 0.1862 & 0 \\ 0.0552 & 0.0724 & 0.0041 & 0 & -0.1488 \\ 0 & 0.0526 & 0 & -0.0663 & 0 \\ -0.0388 & 0.0327 & 0.0040 & -0.0330 & -0.0560 \\ 0 & 0.1549 & 0.0084 & 0 & -0.2383 \\ -0.0277 & 0 & 0.0127 & 0.0405 & 0 \\ 0.1694 & 0.2734 & 0 & 0.3250 & 0.2122 \\ 0.1401 & -0.1367 & -0.3990 & -0.1685 & 0 \\ 0.0106 & 0.1013 & 0 & -0.0728 & 0.1229 \end{bmatrix}. \quad (85)$$

Based on the pattern of zeroes in $\widetilde{\mathbf{B}}_{Sparsed}$, the optimal factor pattern structure can be determined by zeroing out corresponding elements in $\widehat{\mathbf{\Lambda}}_z$. Therefore, the following is obtained:

$$\widehat{\mathbf{\Lambda}}_{zSparsed} = \begin{bmatrix} 0.3957 & 0 & 0 & -0.3321 & 0 \\ 0.4049 & 0 & 0.1007 & 0 & -0.0582 \\ -0.1034 & 0.3989 & 0.4384 & 0.1959 & 0.3028 \\ 0.6635 & -0.0337 & 0 & 0.3846 & 0 \\ 0.7597 & 0 & 0.3186 & 0.1503 & -0.4232 \\ 0.8915 & -0.0584 & 0.4428 & -0.0203 & 0.0204 \\ 0 & 0 & 0 & 0.6619 & 0 \\ 0.8099 & 0.2572 & 0.2690 & 0 & -0.2539 \\ 0 & 0.5586 & 0 & -0.3484 & 0 \\ 0.7731 & 0.3984 & 0.0810 & -0.0228 & -0.1305 \\ 0 & 0.3523 & 0.2148 & 0 & -0.3513 \\ 0.8472 & 0 & 0.3290 & 0.0814 & 0 \\ 0.7888 & 0.3985 & 0 & 0.2335 & 0.1519 \\ 0.8530 & -0.0150 & -0.5168 & -0.0077 & 0 \\ 0.4857 & 0.5714 & 0 & -0.2884 & 0.2954 \end{bmatrix}. \quad (86)$$

Using previously estimated factor scores $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{\Lambda}}_{zSparsed}$, residuals $\boldsymbol{\varepsilon}$ in Equation 2 can be estimated. In matrix form, those estimated residuals are noted as $\underset{(n\times p)}{\widehat{\mathbf{Er}}}$. Then three sets of correlation coefficients are estimated, namely correlations between estimated residuals and estimated factor scores $corr\left(\widehat{\mathbf{Er}}, \widehat{\mathbf{F}}\right)$, correlations between original variables and estimated factor scores $corr\left(\mathbf{X}, \widehat{\mathbf{F}}\right)$, and interfactor correlations $corr\left(\widehat{\mathbf{F}}, \widehat{\mathbf{F}}\right)$.

$$corr\left(\widehat{\mathbf{Er}}, \widehat{\mathbf{F}}\right) = \begin{bmatrix} -0.1614 & 0.3785 & -0.2937 & -0.1107 & 0.2957 \\ -0.1387 & 0.3539 & 0.0042 & 0.2272 & 0.0153 \\ -0.1828 & -0.0993 & -0.0426 & -0.1432 & -0.0256 \\ -0.0561 & -0.0472 & -0.3489 & -0.0903 & 0.4743 \\ -0.0836 & 0.2058 & 0.0619 & -0.1256 & 0.0023 \\ -0.0075 & -0.0101 & -0.0674 & -0.1341 & 0.0401 \\ 0.5719 & -0.2824 & -0.1619 & -0.0854 & 0.0712 \\ 0.0319 & 0.0283 & 0.0077 & -0.3172 & 0.0053 \\ 0.2420 & -0.0444 & -0.0846 & -0.0753 & 0.4788 \\ 0.0399 & 0.0237 & -0.0251 & -0.0660 & 0.0586 \\ 0.8766 & -0.0437 & 0.0439 & -0.0402 & -0.0571 \\ -0.0531 & 0.3552 & -0.0325 & -0.1036 & 0.2370 \\ 0.1395 & 0.0373 & 0.5183 & -0.0671 & 0.1745 \\ -0.0070 & -0.0102 & -0.0677 & -0.1364 & 0.0299 \\ 0.0381 & 0.0045 & -0.0678 & -0.0636 & 0.0986 \end{bmatrix}. \quad (87)$$

As can be seen from Equation 87, most of the correlation coefficients are very close to 0, indicating that the correlations between the residuals and the extracted factors are almost nonexistent. Therefore, the assumption outlined in Equation 9 is satisfied.

$$corr\left(\mathbf{X}, \widehat{\mathbf{F}}\right) = \begin{bmatrix} 0.2460 & 0.3072 & -0.2493 & -0.4766 & 0.2413 \\ 0.2551 & 0.3247 & 0.1105 & 0.2114 & -0.0597 \\ -0.2689 & 0.3597 & 0.4502 & 0.1150 & 0.3453 \\ 0.5739 & -0.1033 & -0.2313 & 0.3406 & 0.3008 \\ 0.7136 & 0.0577 & 0.3258 & 0.1056 & -0.4590 \\ 0.8566 & -0.0992 & 0.4511 & -0.0795 & 0.0261 \\ 0.4020 & -0.2092 & -0.1074 & 0.6716 & 0.0356 \\ 0.7888 & 0.2455 & 0.2749 & -0.1426 & -0.2839 \\ 0.1527 & 0.5531 & -0.0540 & -0.4398 & 0.3499 \\ 0.7400 & 0.3916 & 0.0858 & -0.0765 & -0.1434 \\ 0.7280 & 0.3374 & 0.2216 & -0.0201 & -0.3928 \\ 0.7789 & 0.1584 & 0.3382 & 0.0144 & 0.1111 \\ 0.7574 & 0.3921 & 0.2482 & 0.2071 & 0.1732 \\ 0.7892 & -0.0734 & -0.5134 & -0.0836 & -0.0941 \\ 0.4349 & 0.5761 & 0.0230 & -0.3701 & 0.3341 \end{bmatrix}. \quad (88)$$

As can be seen from Equation 88, the pattern of positive and negative values in this matrix is almost the same as that in the factor loading matrix with the exception of only one matrix element, which is supportive of the effectiveness of the techniques presented in the study.

$$corr\left(\widehat{\mathbf{F}}, \widehat{\mathbf{F}}\right) = \begin{bmatrix} 1.0000 & -0.0535 & 0.0308 & -0.0316 & -0.0622 \\ -0.0535 & 1.0000 & 0.0222 & 0.0010 & 0.0015 \\ 0.0308 & 0.0222 & 1.0000 & 0.0177 & 0.1024 \\ -0.0316 & 0.0010 & 0.0177 & 1.0000 & -0.0236 \\ -0.0622 & 0.0015 & 0.1024 & -0.0236 & 1.0000 \end{bmatrix}. \quad (89)$$

As can be seen from Equation 89, all correlation coefficients are very close to zero, indicating that the extracted factors are statistically independent and orthogonal of each other.

Figure 10 presents the optimal factor pattern structure that has just been identified. All factor loadings in the figure are available in Equation 86.

6.3 Example 3. *Soil Evaporation Data*

In this data set, each of the 46 subjects is measured on 11 variables. So the final data set used for the regression model $\mathbf{F} = \mathbf{X}\,\mathbf{B} + \mathbf{E}$ has a total of 11 variables, or $p = 11$. Background information of the 11 variables is not available.

Next, maximum likelihood factor analysis is run using $MATLAB^{\text{TM}}$'s $factoran$ function. When there are 11 variables, this function can fit up to 6 factors. And as is described above, for each fitted factor model, six information criteria are scored for the purpose of evaluating the model-data fit. Table 6 has in it the information criterion scores for all fitted factor models.

$ICOMP1$ or $ICOMPIFIM$ is minimized at 968.35 when $m = 5$. So the best approximating factor model is selected as the one with 5 factors. As a result, the Genetic Algorithm is run for the 5-factor model.

The GA parameters used for this data set are identical to those for the previous one. Presented in Figure 11, Figure 12, and Figure 13 are the GA progress graphs which are generated under the same rules as in the previous example. In each

graph, the average fitness function value corresponds to the zigzag plot on top whereas the minimum fitness function value the other zigzag plot.

It is evident from these graphs that better models are selected through an evolutionary process. As GA moves from one generation to the next, ICOMPIFIM decreases substantially. Presented in Table 7 is the selected optimal MVR subset model from the three runs of GA. For this subset, ICOMPIFIM is minimized at -1479.9. Note that the columns represent factors and the rows original variables.

When all GA generations are finished, the coefficient matrix $\mathbf{B}$ in $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ is estimated using the $FGLS$ method. In this problem, $p = 11$ and $m = 5$. So $\widetilde{\mathbf{B}}$ should be 11 by 5.

$$\underset{(11\times 5)}{\widetilde{\mathbf{B}}} = \begin{bmatrix} 0.0101 & -0.1379 & 0 & 0 & 0.3618 \\ -0.0026 & -0.0124 & 0.0959 & 0 & 0 \\ 0.0218 & 0.0418 & 0 & -0.0966 & -0.1404 \\ 0 & -0.0075 & 0 & 0 & -0.0131 \\ -0.0169 & -0.0578 & 0.3110 & 0.0976 & -0.1775 \\ 0.0191 & 0.0586 & -0.0783 & 0.0492 & 0.0846 \\ -0.0514 & -0.1977 & -0.0517 & 0.1822 & -0.2808 \\ -0.0147 & -0.0332 & 0 & 0.0705 & -0.0949 \\ -0.0035 & 0.0494 & -0.0256 & -0.0511 & 0.0539 \\ 0.0000 & 0.0001 & 0.0048 & -0.0002 & 0 \\ 0.0008 & 0 & 0 & 0.0056 & 0 \end{bmatrix}. \tag{90}$$

Based on $\widetilde{\mathbf{B}}$, the residuals for the MVR model $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ are computed in the same way as in the previous example, vectorized and plotted in a histogram.

Figure 14 is the histogram of the vectorized residuals.

In Figure 14, the residuals are small, ranging from -1.25 to 1.25. They cluster around 0 and are approximately normally distributed. The histogram shows that the fitted regression model provides a good fit, hence supporting the effectiveness of the new approach to the identification of the best factor pattern structure.

6.4 Example 4. *Gelpo Data*

This data set contains a comparison of 41 countries according to 10 different political and economic parameters. So the final data used for the regression model $\mathbf{F} = \mathbf{X}\ \mathbf{B} + \mathbf{E}$ has a total of 10 variables, or $p = 10$. Background information of the 10 variables is presented in Table 8.

Next, maximum likelihood factor analysis is run using $MATLAB^{\text{TM}}$'s $factoran$ function. When there are 10 variables, this function can fit up to 6 factors. And as is described above, for each fitted factor model, six information criteria are scored for the purpose of evaluating the model-data fit of each factor model. Table 9 has in it the information criterion scores for all fitted factor models.

$ICOMP1$ or $ICOMPIFIM$ is minimized at 959.5 when $m = 3$. So the best approximating factor model is selected as the one with 3 factors. As a result, the

Genetic Algorithm is run for the 3-factor model.

The GA parameters remain unchanged from the previous example. Presented in Figure 15, Figure 16, and Figure 17 are the GA progress graphs which are generated under the same rules as previously. In each graph, the average fitness function value corresponds to the zigzag line on top whereas the minimum fitness function value the other zigzag line.

As the GA generations are being performed, each of the six zigzag lines has a marked tendency to go down, indicating better models are being iteratively identified. These graphs are supportive of the effectiveness of the GA in finding models that provide a better fit. Table 10 has in it the selected optimal MVR subset model from the three runs of GA. For this subset, ICOMPIFIM is minimized at -2107.5. Note that the columns represent factors and the rows original variables.

When all GA generations are finished, the regression weight matrix $\mathbf{B}$ in $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ is estimated using the $FGLS$ method. In this problem, $p = 10$ and $m = 3$. So $\widetilde{\mathbf{B}}$ should be 10 by 3.

$$\underset{(10\times 3)}{\widetilde{\mathbf{B}}} = \begin{bmatrix} -0.0000 & -0.0008 & -0.0001 \\ 0.0001 & -0.0001 & 0.0001 \\ -0.4974 & 0.6160 & 0.9069 \\ 0.0002 & 0.0140 & 0.0006 \\ -0.0008 & -0.0413 & -0.0026 \\ 0.0053 & 0.4018 & 0.0248 \\ 0.0007 & -0.0151 & -0.0283 \\ 0.0001 & 0 & -0.0031 \\ 0.0000 & -0.0007 & -0.0001 \\ 0.0001 & 0.0015 & 0 \end{bmatrix}. \tag{91}$$

Based on $\widetilde{\mathbf{B}}$, the residuals for the MVR model $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ are computed and plotted in the same way as in the previous example. Figure 18 is the histogram of the vectorized residuals.

The vectorized residuals are small, ranging from -0.8 to 0.5. They cluster around 0 and are only a little bit skewed. The histogram shows that the fitted MVR model based on $\widetilde{\mathbf{B}}$ provides a good fit, hence supporting the effectiveness of the new approach to the identification of the best factor pattern structure.

6.5 Example 5. *Medical School Test Data*

The last example is based on a data set consisting of 142 rows and 24 columns. Each element in this 142 by 24 matrix is a score from a new medical school student on an item of a psychological test. A total of 142 medical school students are involved. Background information of the 24 test items is not available.

Next, maximum likelihood factor analysis is run using $MATLAB^{TM}$'s $factoran$ function. When there are 24 variables, this function can fit up to 17 factors. And as is described above, for each fitted factor model, six information criteria are scored for the purpose of evaluating the model-data fit. Table 11 has in it the information criterion scores for the first 12 factor models.

$ICOMPC1$ or $ICOMPIFIM$ is minimized at 6994.6 when $m = 9$. So the best approximating factor model is selected as the one with 9 factors. As a result, the Genetic Algorithm is run for the 9-factor model.

The GA parameters used for this data set are identical to those for the previous one. Presented in Figure 19, Figure 20, and Figure 21 are the GA progress graphs which are generated under the same rules as previously. In each graph, the average fitness function value corresponds to the zigzag plot on top whereas the minimum fitness function value the other zigzag plot.

Like in the previous examples, those GA progress graphs show the effectiveness of the GA in finding better models through an iterative process. Each of the six zigzag lines drops substantially with each GA generation, hence the identification of a better model. Presented in Table 12 and Table 13 is the selected optimal MVR subset model from the three runs of GA. For this subset, ICOMPIFIM is

minimized at -3157.4. Note that the columns represent factors and the rows original variables.

When all GA generations are finished, the regression coefficient matrix $\mathbf{B}$ in $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ is estimated using the $FGLS$ method. In this problem, $p = 24$ and $m = 9$. So $\widetilde{\mathbf{B}}$ should be 24 by 9. The estimated matrix $\widetilde{\mathbf{B}}$ is omitted here to save space.

Based on $\widetilde{\mathbf{B}}$, the residuals for the model $\mathbf{F} = \mathbf{XB} + \mathbf{E}$ are computed and plotted in the same way as in the previous example. Figure 22 has in it the histogram of the vectorized residuals.

The histogram shows that the residuals from the fitted MVR model based on $\widetilde{\mathbf{B}}$ are small, ranging from -1.5 to 1.5. In addition, the residuals cluster around 0 and are approximately normally distributed. The histogram indicates the fitted model provides a good fit, hence supporting the effectiveness of the new approach to the identification of the best factor pattern structure.

7 Conclusions

The study presents a new approach to expert model selection in maximum likelihood factor analysis using information criteria. The study emphasizes the use of Bozdogan's $ICOMP-type\ criteria$: $ICOMP\,(IFIM)$, $ICOMPIFIM_{PEU\_Mis}$, $ICOMPIFIM_{PEU\_Mis\_LN}$ (Bearse & Bozdogan, 2000; Bozdogan 2000) and it compares factor model selection results using $ICOMP\ criteria$ with those from other well-established model selection criteria: AIC (Akaike, 1973, 1987), CAIC (Bozdogan & Ramirez, 1987), and SBC (Schwartz, 1978). At the same time, $ICOMPIFIM$ is also used as the fitness function in the implementation of the Genetic Algorithm for the purpose of selecting the optimal MVR subset. Based on model selection results from the GA, the study finds the extent to which extracted factors depend on the original data by building a multivariate regression model which relates factor scores to the original data set.

The study removes the subjectivity in the selection of factor models. One traditional approach to the determination of the number of factors counts on eyeballing the scree plot of the eigenvalues. This causes plenty of subjectivity

in model selection because different people may draw different conclusions about at which eigenvalue the scree plot tends to level off and, thus, come up with different answers to the number of factors that need to be extracted. The use of information criteria eliminates the subjectivity in that the best model is selected as the one that minimizes information criteria. Stated differently, information criteria map the performance of a candidate model to a scalar value on which to base subsequent conclusions on model-data fit.

The study implements the Genetic Algorithm (Bearse & Bozdogan, 2000; Holland, 1975) in the determination of the best MVR subset. The GA provides a computationally inexpensive approach to best MVR subset selection and this study is supportive of its use in complex modeling situations where the traditional all-possible-subset technique fails to work. This study provides evidence to show how the GA works smartly to quickly find an optimal MVR subset through an iterative and evolutionary process after the required parameters are given. Since the specification of such GA parameters as population size, number of generations, probability of crossover, probability of mutation and the fitness function of choice is vital to the performance of the algorithm, further study on this topic is warranted. However, this study does support the use of $ICOMPIFIM$ which

performs well as the GA objective/fitness function.

The study builds a multivariate regression model that predicts factor scores based on original variables. The GA simultaneously takes into account all response variables in determining the champion MVR model structure, or the optimal subset of original variables used for the prediction of each response or each of the factor score variables. The estimation of the regression coefficients is based on a method known as *feasible generalized least squares* ($FGLS$). The marriage of the GA and the FGLS method proves to be effective as the residuals from each fitted MVR model appear to be small and normally distributed, and they also tend to cluster around 0.

In conclusion, the new approach presented in the study successfully unifies EFA and CFA by providing regular EFA with a computationally efficient and effective algorithm for optimal factor pattern search. This approach is recommended for use in similar modeling situations.

Appendix A: List of References

REFERENCES

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Budapest, Hungary: Akademiai Kiado.

[2] Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317-332.

[3] Anderson, T. W., & Rubin, H. (1956). Statistical inference in factor analysis. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 5,* 111-150.

[4] Bearse, P. M., & Bozdogan, H. (2000, May). *Multivariate regressions, genetic algorithms, and information complexity: A three-way hybrid.* Paper presented at the International Conference on Measurement and Multivariate Analysis (ICMMA) and Dual Scaling, Banff, Canada.

[5] Bensmail, H., & Bozdogan, H. (2002). Regularized kernel discriminant analysis with optimally scaled data. In S. Nishisato, Y. Baba, H. Bozdogan &

K. Kanefuji (Eds.), *Measurement and Multivariate Analysis* (pp. 133-144). Tokyo, Japan: Springer Verlag.

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

[7] Bozdogan, H. (1996). *A new informational complexity criterion for model selection: The general theory and its applications.* Invited paper presented at the annual meeting of the Institute for Operations Research & the Management Sciences (INFORMS), Washington, DC.

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

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

[10] Bozdogan, H., & Ramirez, D. E. (1987). An expert model selection approach to determine the best pattern structure in factor analysis models. In H. Bozdogan & A. K. Gupta (Eds.), *Multivariate Statistical Modeling and Data Analysis* (pp. 35-60). Dordrecht, Netherlands: Reidel Publishing Company.

[11] Bozdogan, H., & Shigemasu, K. (1998). *Bayesian factor analysis model and choosing the number of factors using a new informational complexity criterion*. Invited paper presented at the annual meeting of the International Federation of Classification Societies (IFCS), Rome, Italy.

[12] Child, D. (1990). *The essentials of factor analysis* (2nd ed.). London, UK: Cassel Educational Limited.

[13] Everitt, B. S. (1984). *An introduction to latent variable models*. London, UK: Chapman & Hall/CRC.

[14] Garson, G. D. (2008, March). *Factor analysis*. Retrieved July 26, 2008, from http://www2.chass.ncsu.edu/garson/pa765/factor.htm

[15] Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago, IL: University of Chicago Press.

[16] Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor, MI: University of Michigan Press.

[17] Johnson, R. A., & Wichern, D. W. (1982). *Applied multivariate statistical analysis*. Upper Saddle River, NJ: Prentice Hall.

[18] Lanning, J. M. (2008, August). *On information criteria and data mining visualization using kernelized regression with GA and decision tree*. Unpublished doctoral dissertation, University of Tennessee, Knoxville.

[19] Lawley, D. N. (1940). The estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh, A-60,* 64-82.

[20] Lawley, D. N. (1942). Further investigations in factor estimation. *Proceedings of the Royal Society of Edinburgh, A-61,* 176-185.

[21] Lawley, D. N., & Maxwell, A. E. (1971). *Factor analysis as a statistical method*. New York, NY: American Elsevier.

[22] Liu, Y. (2007, August). *On robust model selection with information complexity and genetic algorithms*. Unpublished doctoral dissertation, University of Tennessee, Knoxville.

[23] Liu, M., & Bozdogan, H. (2008). Multivariate regression models with power exponential random errors and subset selection using genetic algorithms with information complexity. *European Journal of Pure and Applied Mathematics, 1*(1), 4-37.

[24] Liu, Z., & Bozdogan, H. (2004). Kernel PCA for feature extraction with information complexity. In H. Bozdogan (Ed.), *Statistical Data Mining & Knowledge Discovery* (pp. 309-322). London, UK: Chapman & Hall/CRC.

[25] Lopes, H. F., & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica, 14,* 41-67.

[26] Munro, B. H. (2004). *Statistical methods for health care research* (5th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

[27] Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461-464.

[28] Suhr, D. D. (n.d.). *Exploratory or confirmatory factor analysis.* Retrieved July 26, 2008, from http://www2.sas.com/proceedings/sugi31/200-31.pdf

[29] Zhang, R. (2007, August). *On model selection techniques for kernel-based regression analysis using information complexity measure and genetic algorithms.* Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Appendix B: List of Tables

Table 1: Criterion Scores for Fitted Factor Models for the Sim Data

	1	2	3	4	5	6	7
$ICOMP1$	3054.0	2867.4	**2785.1**	2788.5	2793.0	2795.7	2807.2
$ICOMP2$	3118.9	2978.5	**2954.9**	3034.0	3137.4	3271.6	3459.4
$ICOMP3$	3204.0	**3117.1**	3126.8	3222.8	3340.0	3486.9	3692.0
AIC	3036.6	2831.0	**2743.1**	2751.6	2761.4	2768.4	2778.6
$CAIC$	3123.2	2957.2	**2905.4**	2946.3	2985.0	3017.2	3049.0
SBC	3099.2	2922.2	**2860.4**	2892.3	2923.0	2948.2	2974.0

Table 2: GA MVR Subset for the Sim Data

	f_1	f_2	f_3
x_1	1	1	1
x_2	1	0	1
x_3	1	1	1
x_4	1	1	1
x_5	1	1	1
x_6	1	1	0
x_7	1	1	1
x_8	1	1	1
x_9	1	0	1
x_{10}	1	1	1
x_{11}	1	0	1
x_{12}	1	1	1

Table 3: Variable Description for the Job Data

x_1 :	Form of application letter
x_2 :	Appearance
x_3 :	Academic ability
x_4 :	Likeability
x_5 :	Self-confidence
x_6 :	Lucidity
x_7 :	Honesty
x_8 :	Salesmanship
x_9 :	Experience
x_{10} :	Drive
x_{11} :	Ambition
x_{12} :	Grasp
x_{13} :	Potential
x_{14} :	Keenness to join
x_{15} :	Suitability

Table 4: Criterion Scores for Fitted Factor Models for the Job Data

	1	2	3	4	5
$ICOMP1$	1780.2	1739.3	1709.6	1691.6	**1689.1**
$ICOMP2$	1960.2	3851.3	**1212.2**	1403.6	1463.2
$ICOMP3$	2128.4	4058.8	**1456.5**	1681.6	1766.1
AIC	1660.5	1605.4	1562.5	1532.4	1525.3
$CAIC$	1746.7	1731.7	**1726.1**	1730.5	1755.0
SBC	1716.7	1687.7	1669.1	**1661.5**	1675.0
	6	7	8	9	10
$ICOMP1$	1689.9	1690.7	1698.3	1714.4	1726.1
$ICOMP2$	1493.6	1511.4	1529.9	1553.5	1570.4
$ICOMP3$	1819.8	1864.9	1905.5	1949.4	1979.3
AIC	1521.2	**1510.8**	1511.0	1519.1	1529.1
$CAIC$	1779.7	1795.1	1818.2	1846.5	1873.6
SBC	1689.7	1696.1	1711.2	1732.5	1753.6

Table 5: GA MVR Subset for the Job Data

	f_1	f_2	f_3	f_4	f_5
x_1	1	0	1	1	0
x_2	1	0	1	0	1
x_3	1	1	1	1	1
x_4	1	1	0	1	0
x_5	1	0	1	1	1
x_6	1	1	1	1	1
x_7	0	0	1	1	0
x_8	1	1	1	0	1
x_9	0	1	1	1	0
x_{10}	1	1	1	1	1
x_{11}	0	1	1	0	1
x_{12}	1	0	1	1	0
x_{13}	1	1	1	1	1
x_{14}	1	1	1	1	0
x_{15}	1	1	1	1	1

Table 6: Criterion Scores for Fitted Factor Models for the Soil Data

	1	2	3	4	5	6
$ICOMP1$	1129.0	1012.1	976.99	974.88	**968.35**	978.61
$ICOMP2$	1221.0	1257.4	2234.3	**73.284**	539.02	661.72
$ICOMP3$	1351.5	1449.2	2462.0	**325.06**	815.5	957.77
AIC	1030.2	866.33	809.99	797.52	**777.97**	778.83
$CAIC$	1092.4	956.84	**925.96**	936.12	936.37	954.2
SBC	1070.4	924.84	884.96	887.12	**880.37**	892.2

Table 7: GA MVR Subset for the Soil Data

	f_1	f_2	f_3	f_4	f_5
x_1	1	1	0	0	1
x_2	1	1	1	0	0
x_3	1	1	0	1	1
x_4	0	1	0	0	1
x_5	1	1	1	1	1
x_6	1	1	1	1	1
x_7	1	1	1	1	1
x_8	1	1	0	1	1
x_9	1	1	1	1	1
x_{10}	1	1	1	1	0
x_{11}	1	0	0	1	0

Table 8: Variable Description for the Gelpo Data

x_1 :	Population
x_2 :	Gross internal product per habitant
x_3 :	Rate of increase of the population
x_4 :	Rate of urban population
x_5 :	Rate of illiterate in the population
x_6 :	Rate of students in the population
x_7 :	Expected life time of people
x_8 :	Rate of nutritional needs realized
x_9 :	Number of newspapers & magazines per 1000 habitants
x_{10} :	Number of TV sets per 1000 habitants

Table 9: Criterion Scores for Fitted Factor Models for the Gelpo Data

	1	2	3	4	5	6
$ICOMP1$	988.64	967.03	**959.5**	959.78	964.74	969.88
$ICOMP2$	1075.0	1204.8	2476.5	**238.18**	592.01	688.01
$ICOMP3$	1153.4	1300.7	2586.8	**360.16**	725.83	832.09
AIC	937.08	913.13	**904.81**	905.42	908.55	911.72
$CAIC$	**991.35**	991.82	1005.2	1024.8	1044.2	1061.0
SBC	971.35	**962.82**	968.21	980.82	994.23	1006.0

Table 10: GA MVR Subset for the Gelpo Data

	f_1	f_2	f_3
x_1	0	1	1
x_2	1	1	1
x_3	1	1	1
x_4	1	1	1
x_5	1	1	1
x_6	1	1	0
x_7	1	1	1
x_8	0	1	1
x_9	1	1	1
x_{10}	1	1	1

Table 11: Criterion Scores for Fitted Factor Models for the Test Data

	1	2	3	4	5	6
$ICOMP1$	8771.3	7830.4	7287.9	7150.3	7095.7	7045.3
$ICOMP2$	8919.4	8122.6	7849.9	8395.6	13438	**3702.9**
$ICOMP3$	9270.6	8838.9	8799.1	9440.1	14551	**4883.5**
AIC	8629.6	7487.8	6831.6	6671.6	6610.9	6552.5
$CAIC$	8819.5	7768.6	7199.5	**7122.5**	7141.0	7157.7
SBC	8771.5	7697.6	7106.5	7008.5	7007.0	**7004.7**
	7	8	9	10	11	12
$ICOMP1$	7014.9	6997.5	**6994.6**	7000.8	6998.8	7000.9
$ICOMP2$	5448.4	5885.2	6089.4	6213.5	6287.2	6341.8
$ICOMP3$	6691.9	7184.4	7437.3	7618.5	7742.9	7847.7
AIC	6515.6	6494.4	6490.6	6488.1	6479.8	**6474.0**
$CAIC$	7192.0	7238.1	7297.6	7354.4	7401.5	7447.1
SBC	7021.0	7050.1	7093.6	7135.4	7168.5	7201.1

Table 12: GA MVR Subset for the Test Data - Part 1

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
x_1	1	0	1	0	0	0	1	1	1
x_2	0	1	1	1	0	0	1	1	1
x_3	1	1	1	1	1	0	1	1	0
x_4	1	0	0	1	1	1	1	1	1
x_5	0	0	1	0	1	1	1	0	1
x_6	0	1	0	1	1	0	0	1	0
x_7	1	1	1	1	0	1	0	0	1
x_8	0	1	1	1	0	0	0	0	1
x_9	0	1	1	1	1	1	1	1	1
x_{10}	0	0	1	1	1	1	1	0	0
x_{11}	1	0	1	1	1	0	0	0	1
x_{12}	0	1	1	1	1	1	1	1	0

Table 13: GA MVR Subset for the Test Data - Part 2

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
x_{13}	1	1	1	1	1	0	1	1	1
x_{14}	1	0	0	1	0	1	1	1	0
x_{15}	0	1	0	1	0	1	1	0	1
x_{16}	0	1	0	1	1	0	0	1	1
x_{17}	1	0	1	1	1	1	0	0	1
x_{18}	1	1	1	1	1	1	1	1	1
x_{19}	0	0	1	1	1	1	0	1	1
x_{20}	0	1	0	1	0	1	1	0	1
x_{21}	1	1	0	1	1	1	0	0	0
x_{22}	0	0	1	0	0	0	1	1	1
x_{23}	0	1	1	1	0	1	1	1	0
x_{24}	0	1	0	1	1	0	0	1	1

Appendix C: List of Figures

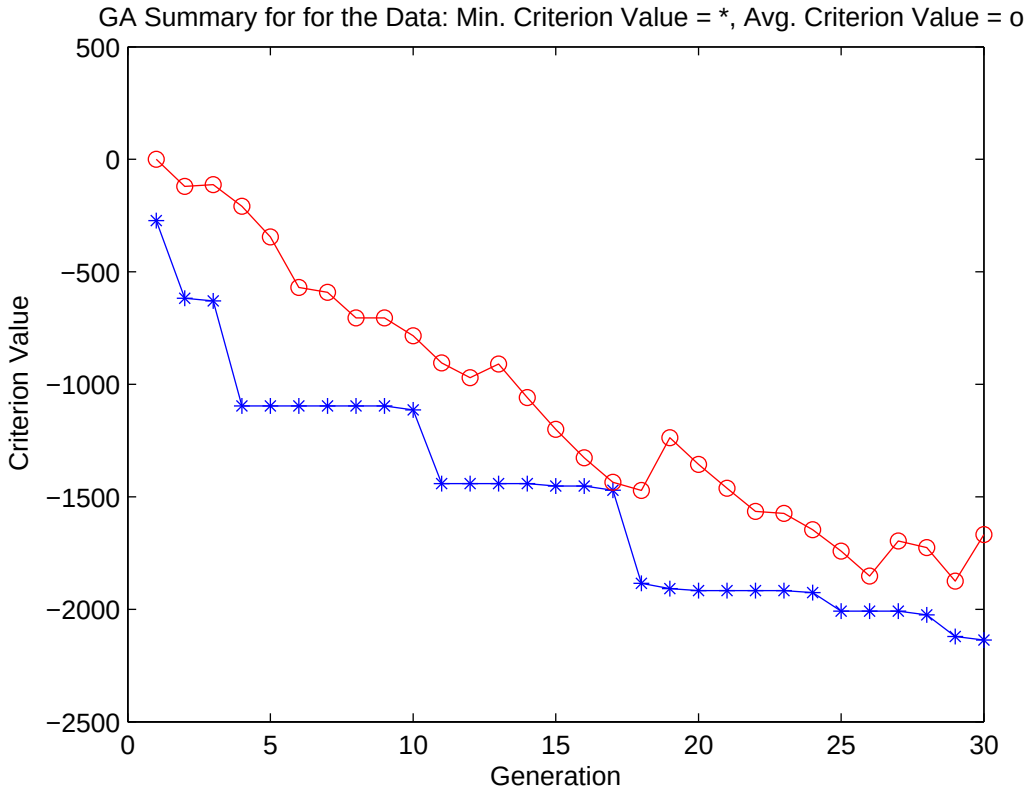

Figure 1: First GA Run for the Sim Data

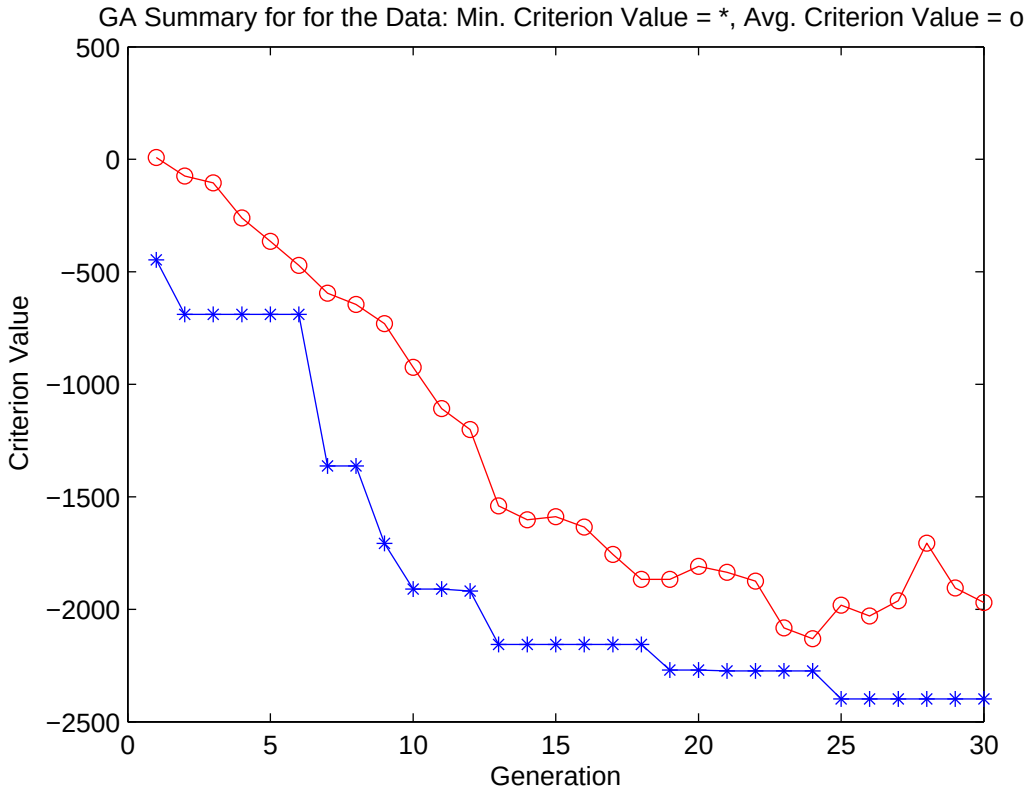

Figure 2: Second GA Run for the Sim Data

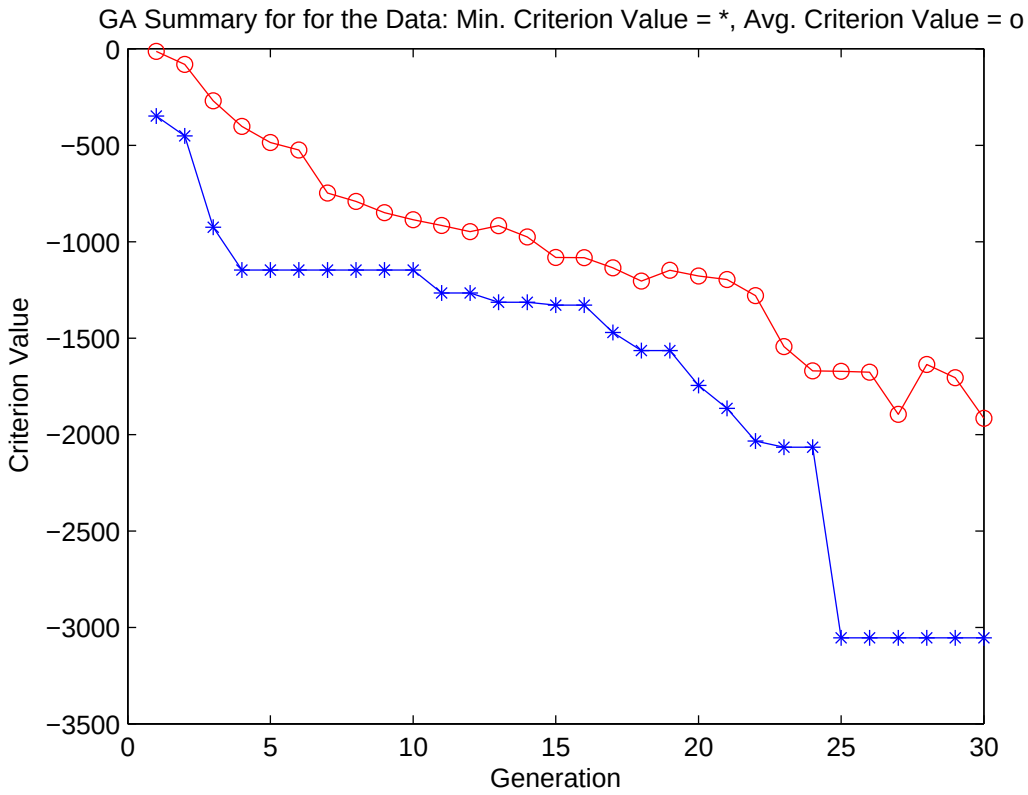

Figure 3: Third GA Run for the Sim Data

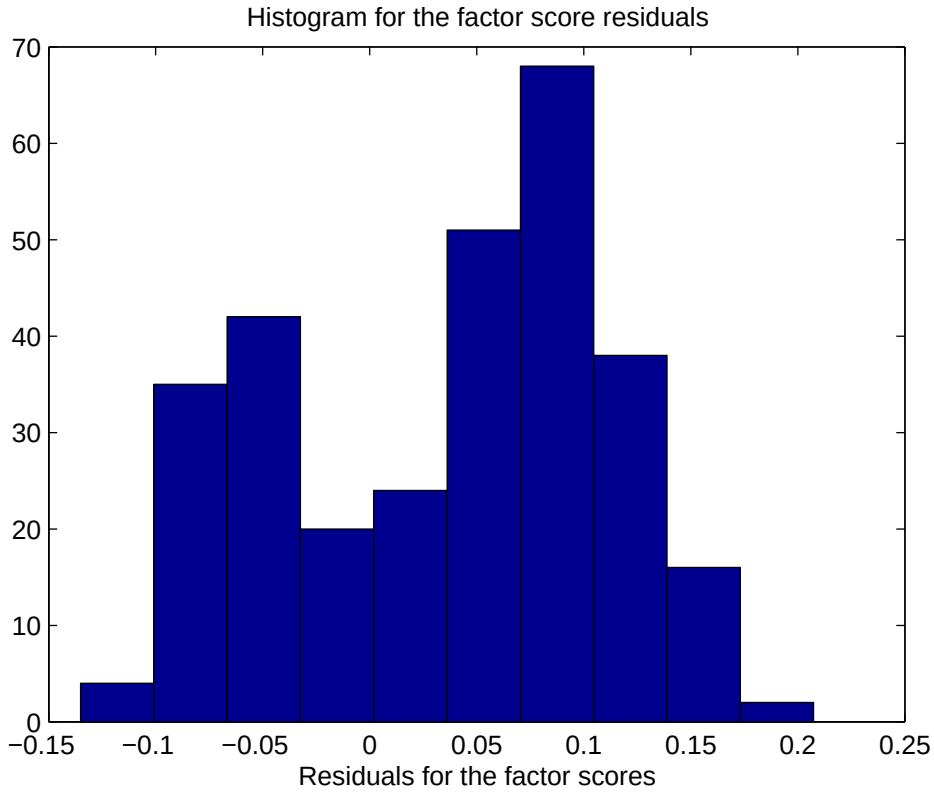

Figure 4: Distribution of Factor Score Residuals for the Sim Data

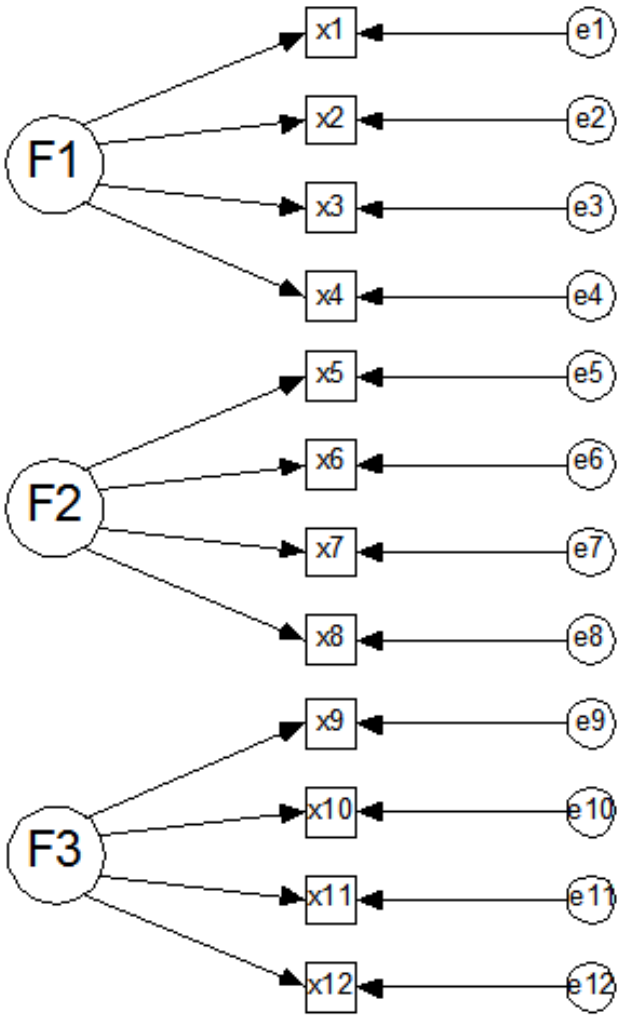

Figure 5: Factor Pattern Diagram for the Sim Data

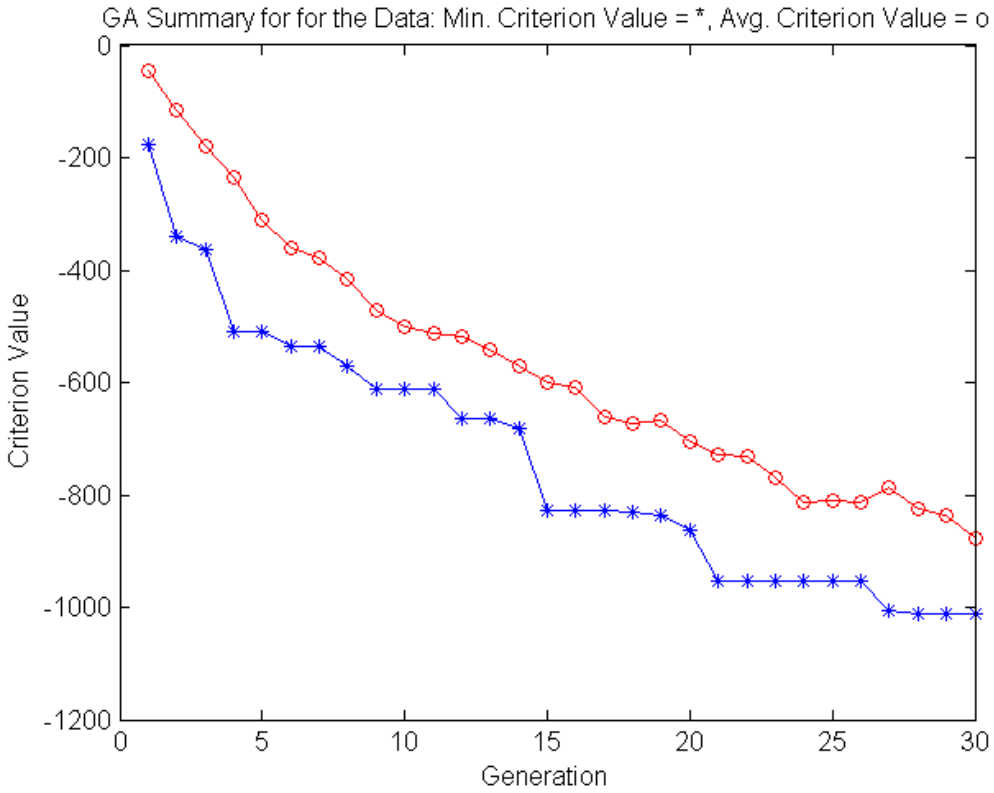

Figure 6: First GA Run for the Job Data

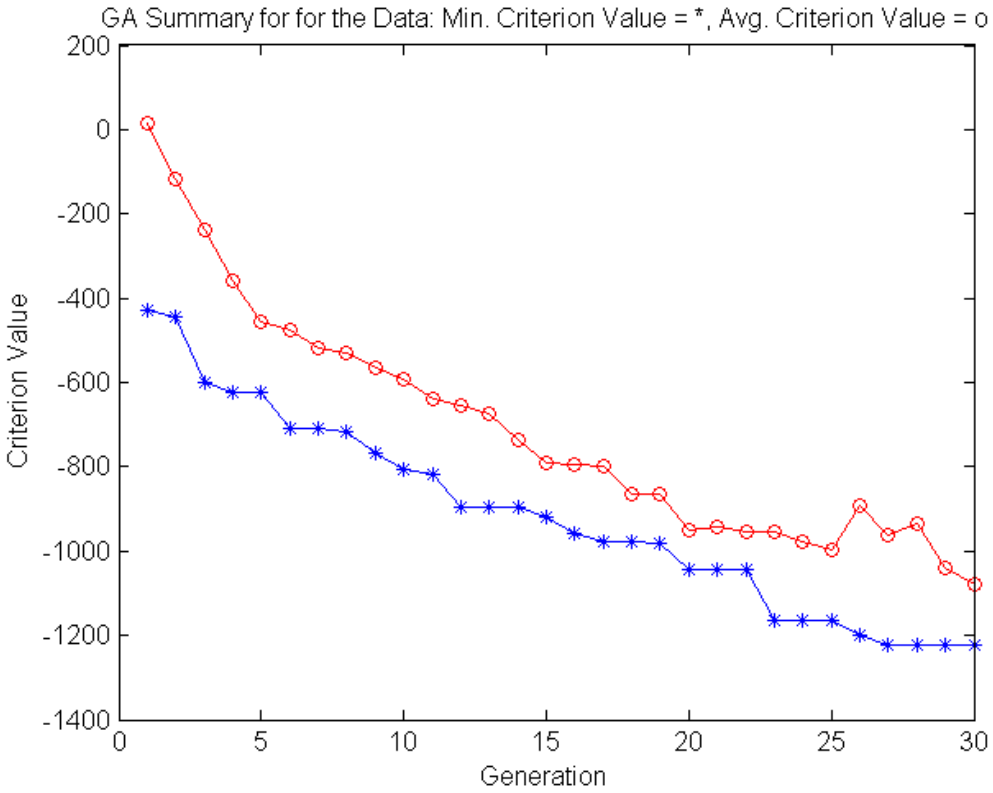

Figure 7: Second GA Run for the Job Data

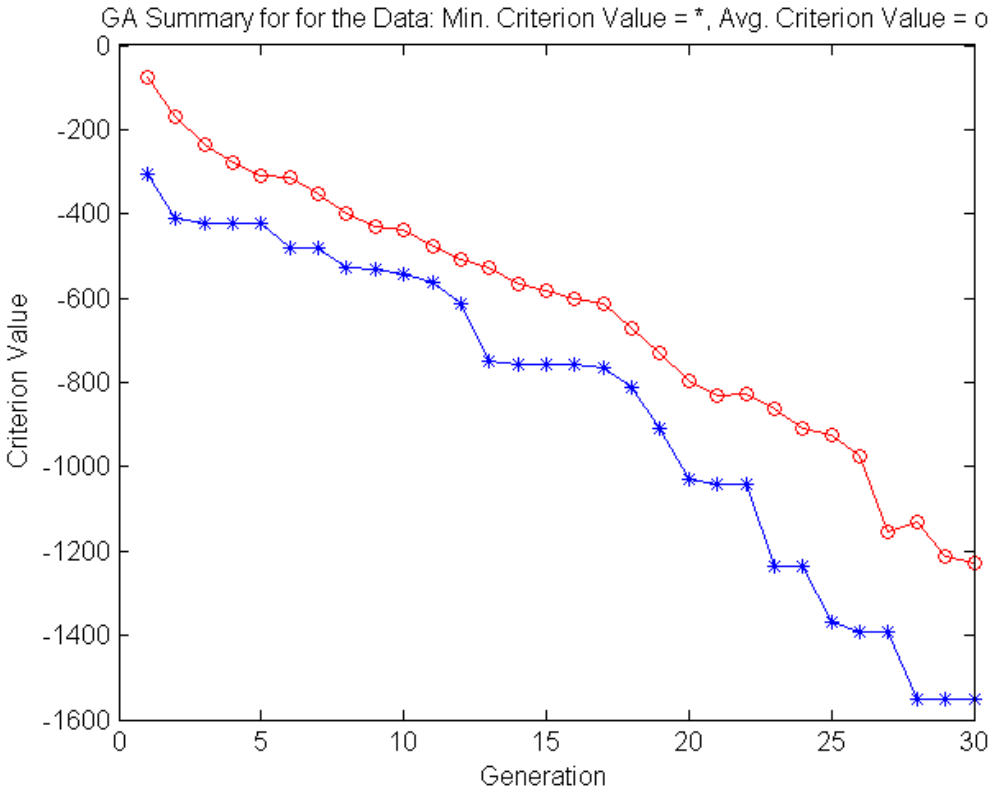

Figure 8: Third GA Run for the Job Data

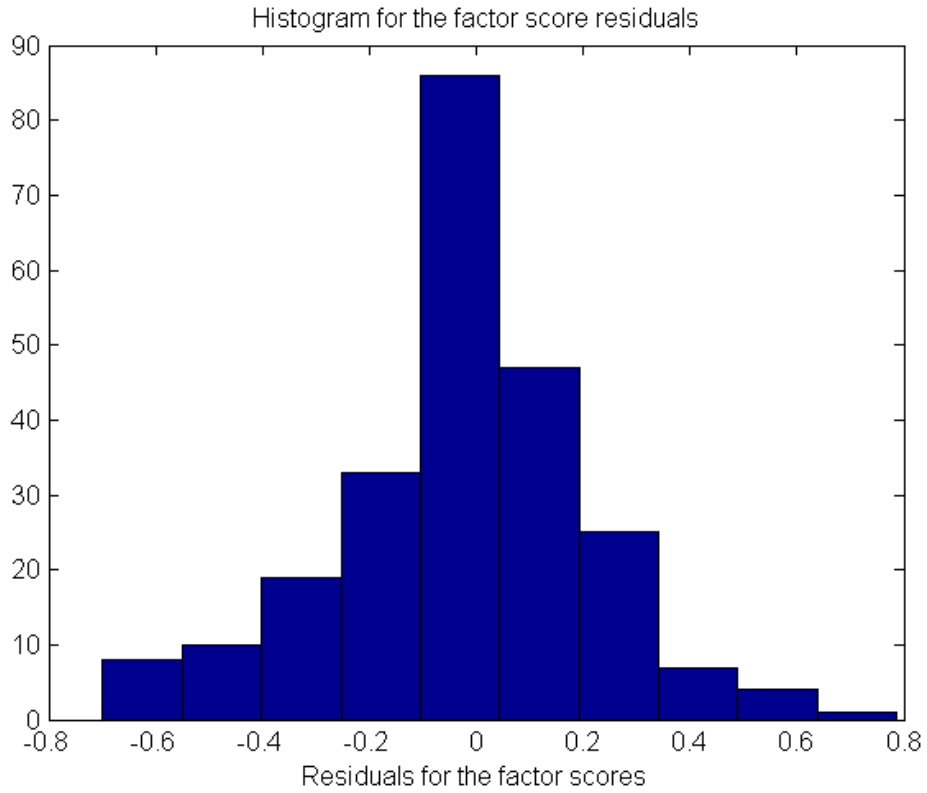

Figure 9: Distribution of Factor Score Residuals for the Job Data

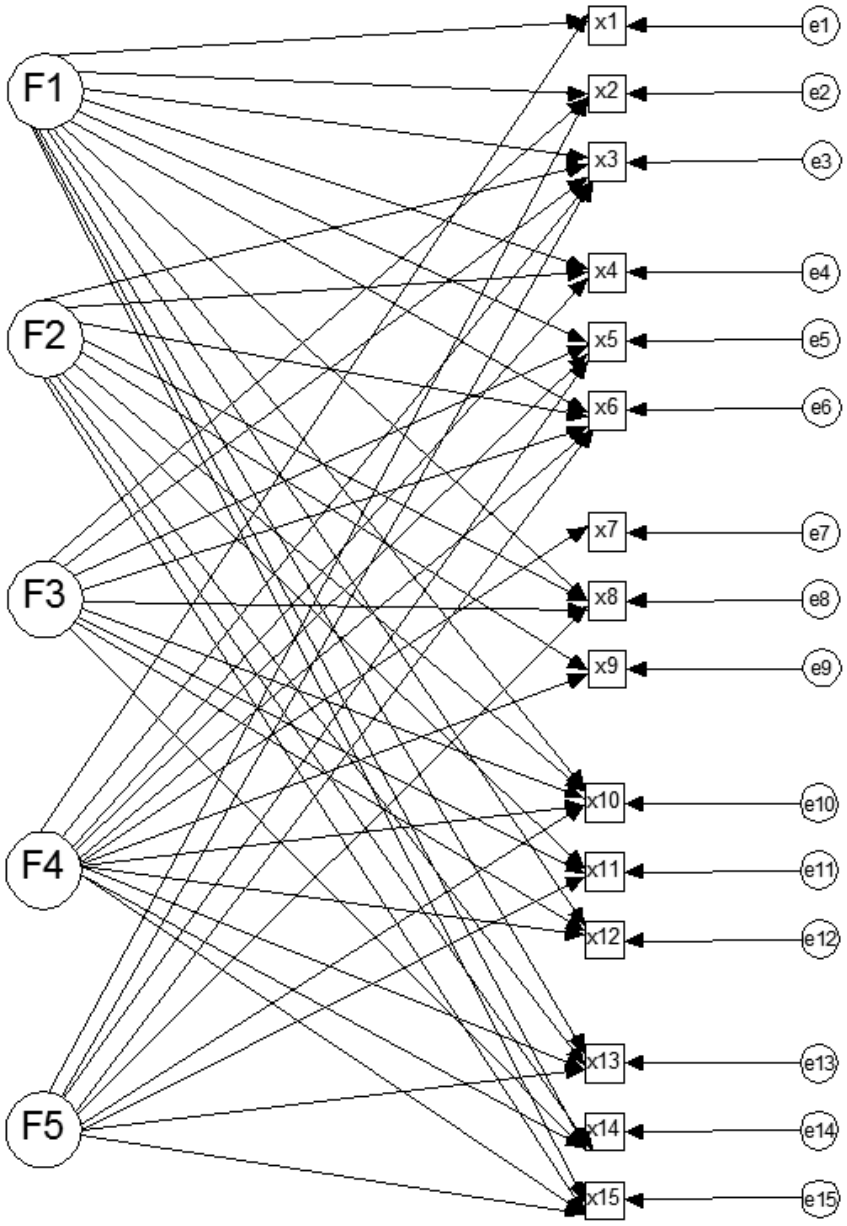

Figure 10: Factor Pattern Diagram for the Job Data

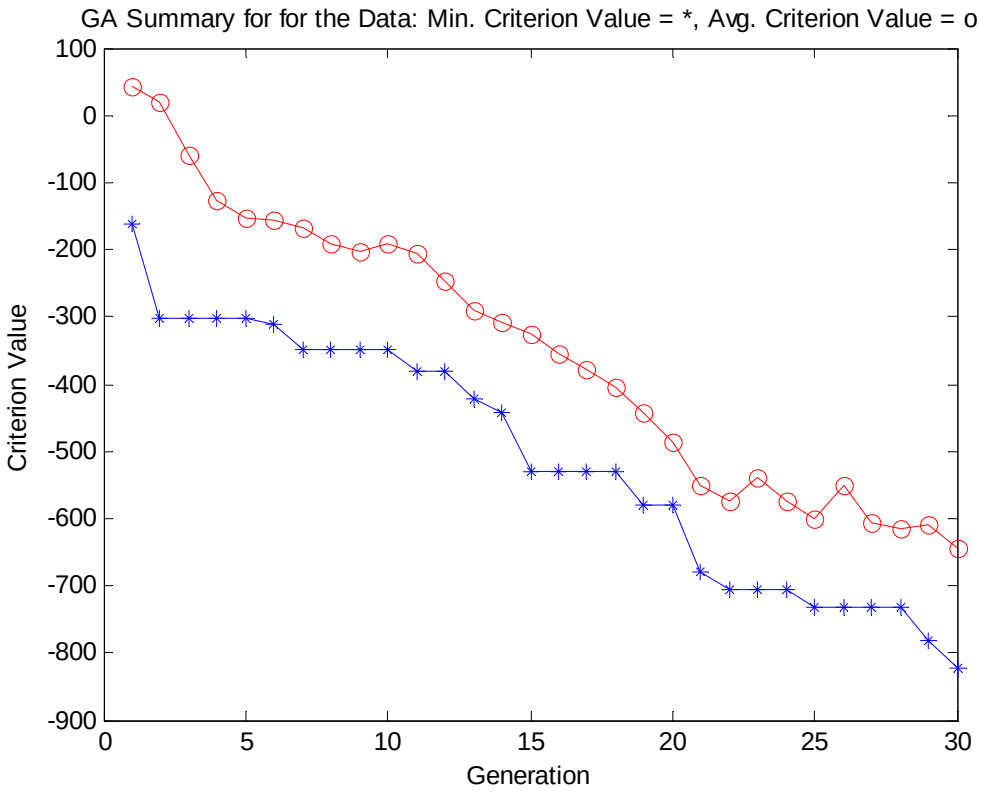

Figure 11: First GA Run for the Soil Data

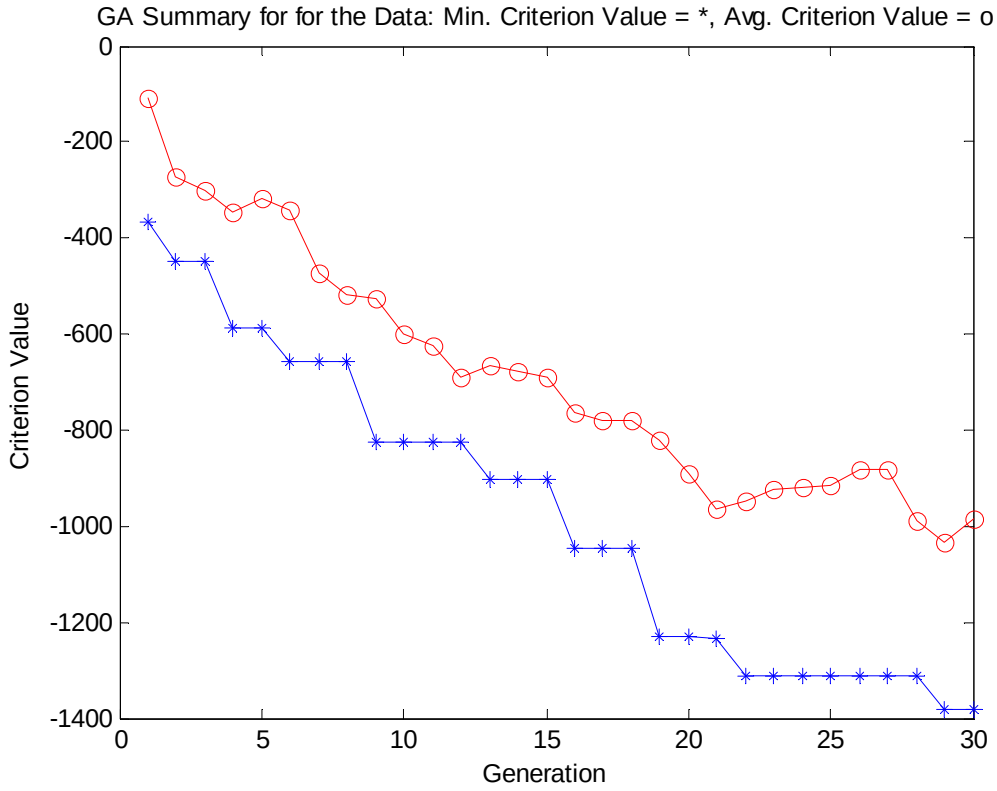

Figure 12: Second GA Run for the Soil Data

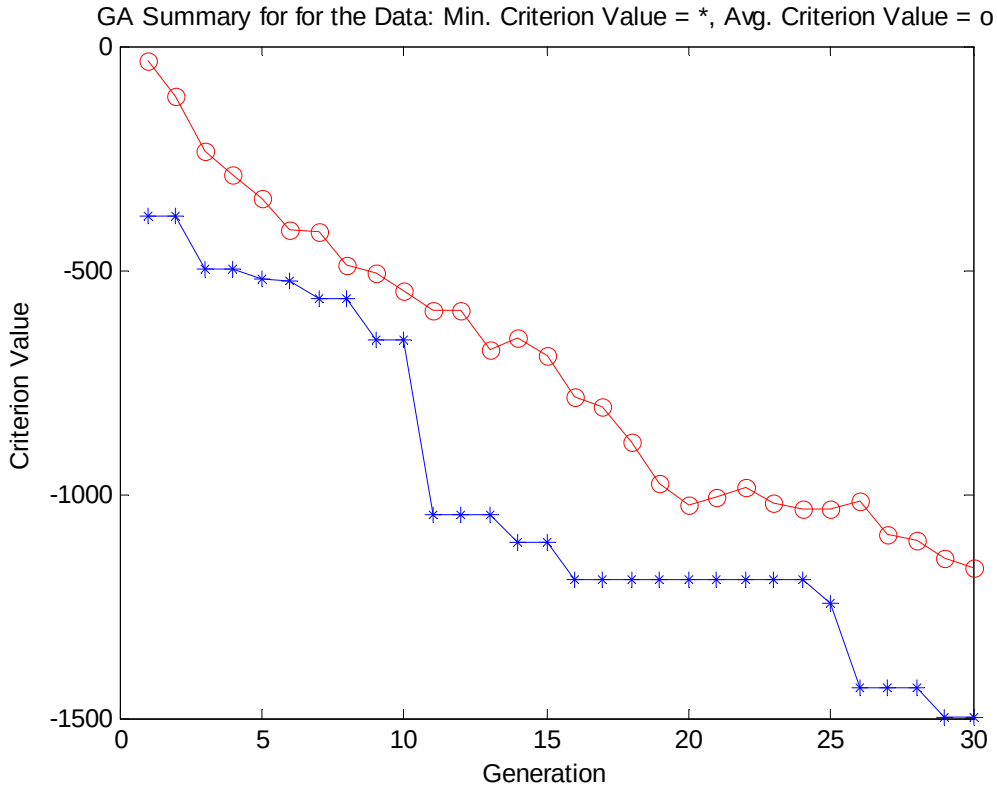

Figure 13: Third GA Run for the Soil Data

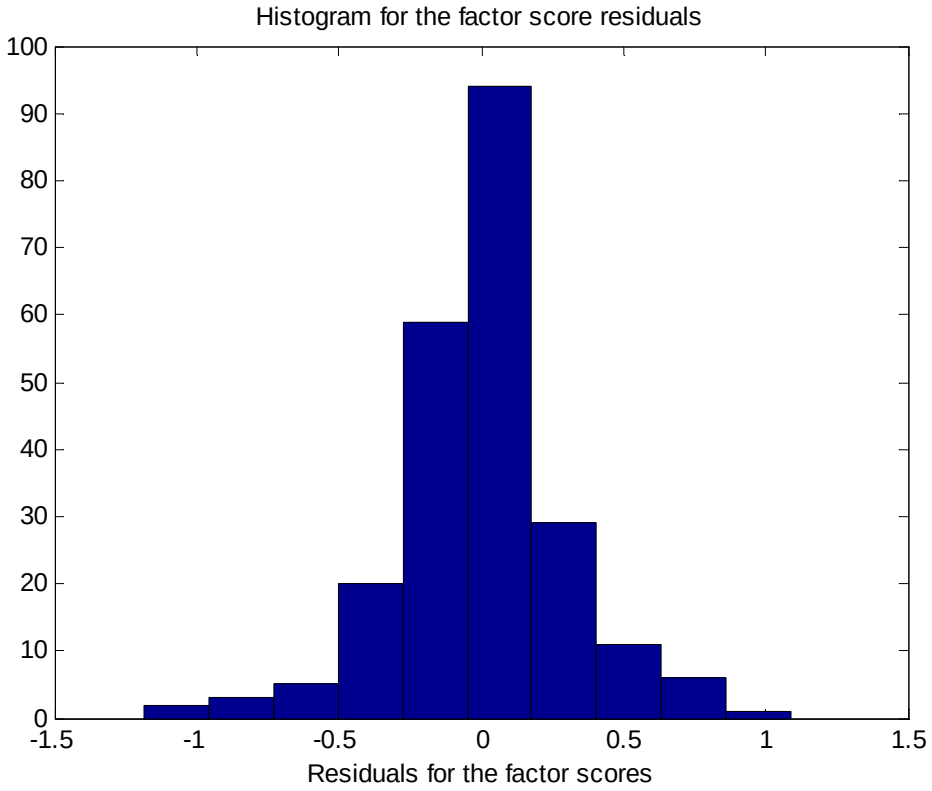

Figure 14: Distribution of Factor Score Residuals for the Soil Data

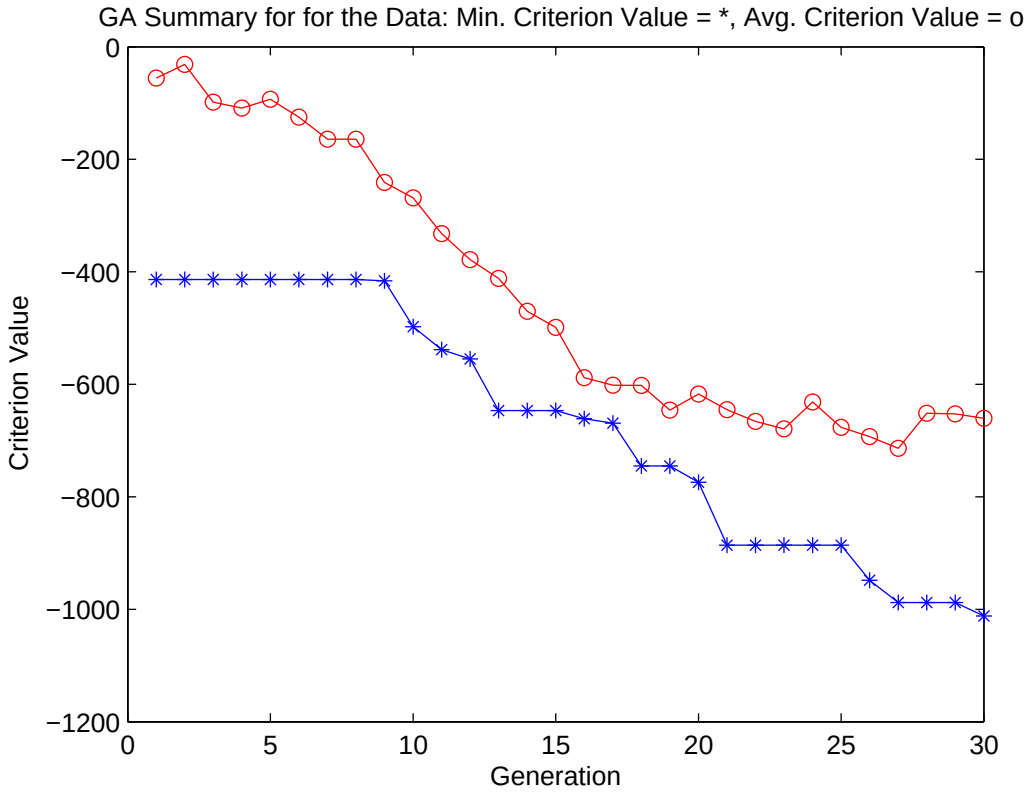

Figure 15: First GA Run for the Gelpo Data

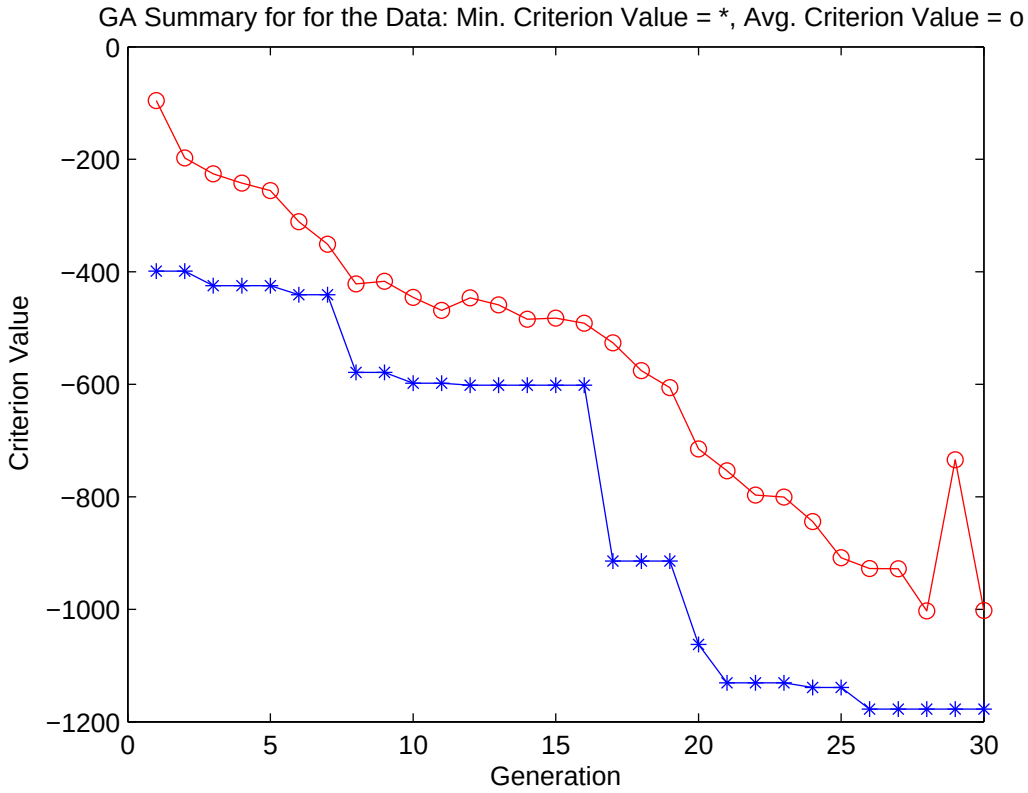

Figure 16: Second GA Run for the Gelpo Data

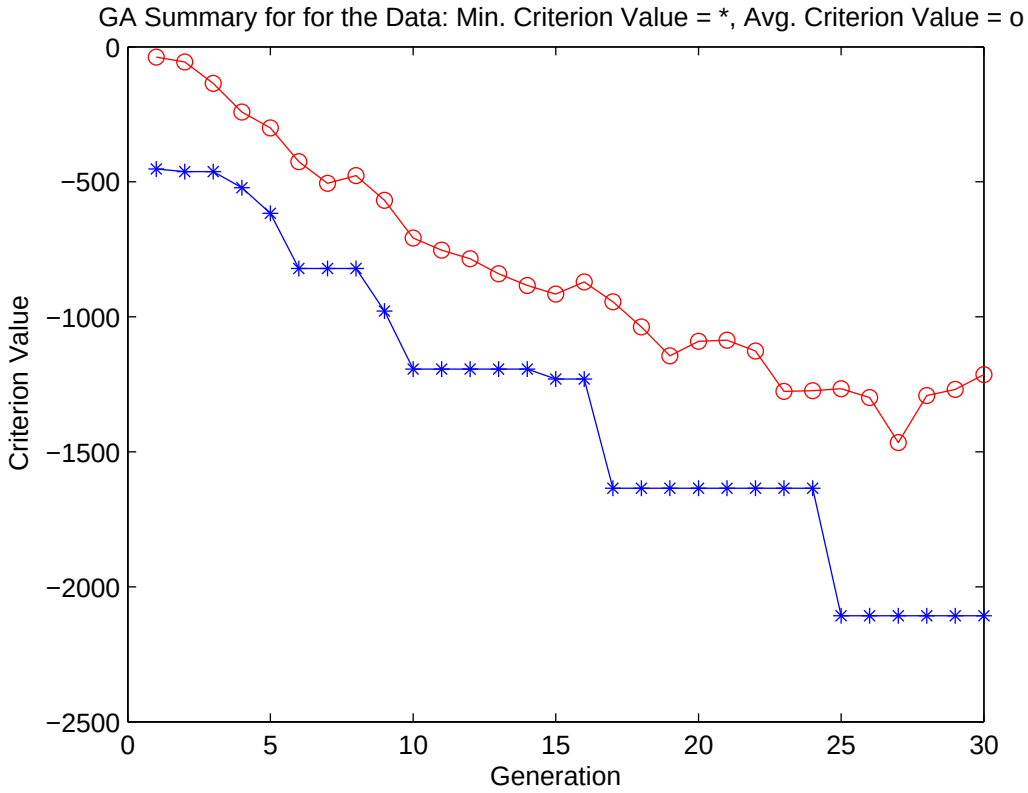

Figure 17: Third GA Run for the Gelpo Data

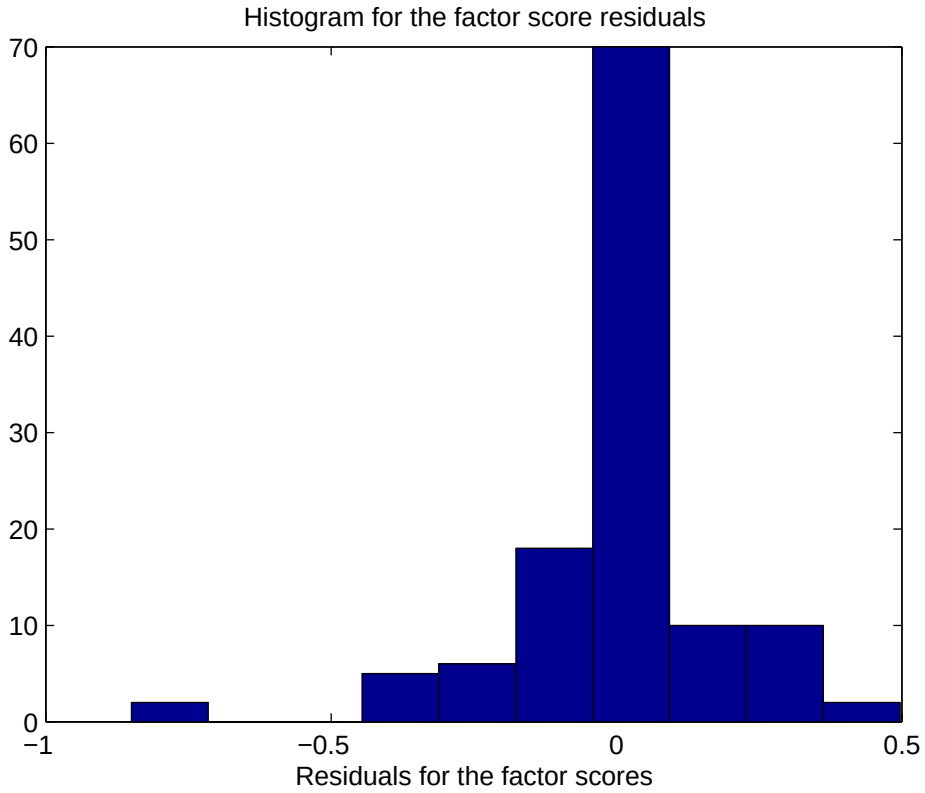

Figure 18: Distribution of Factor Score Residuals for the Gelpo Data

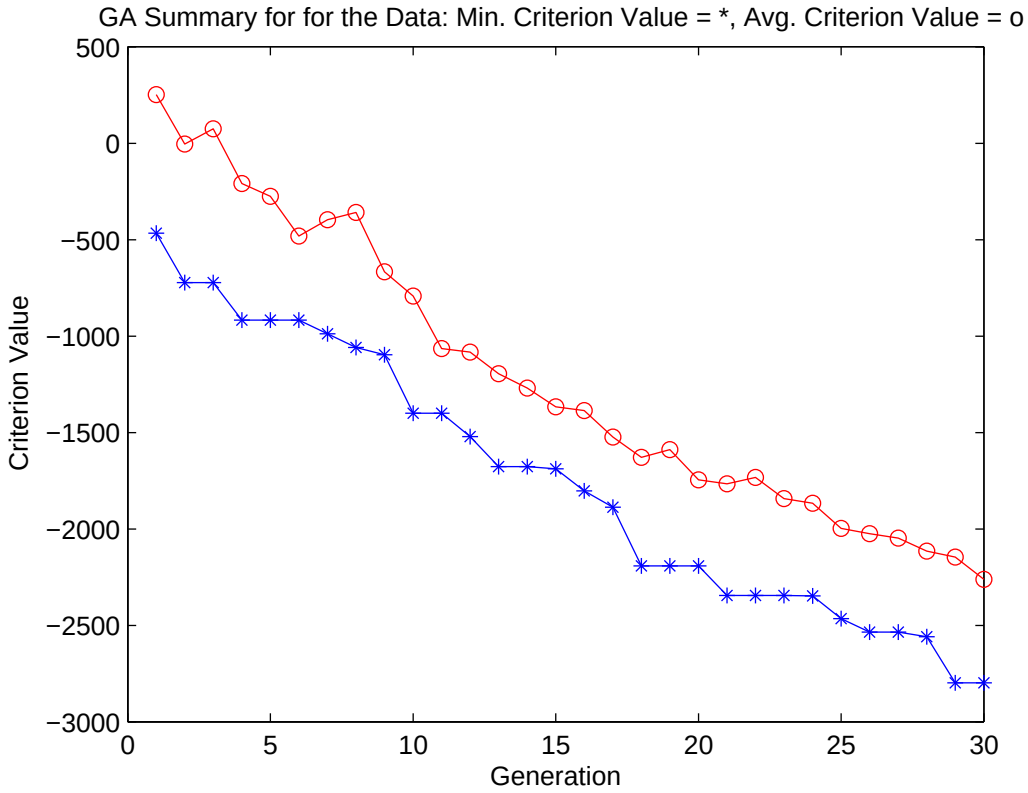

Figure 19: First GA Run for the Test Data

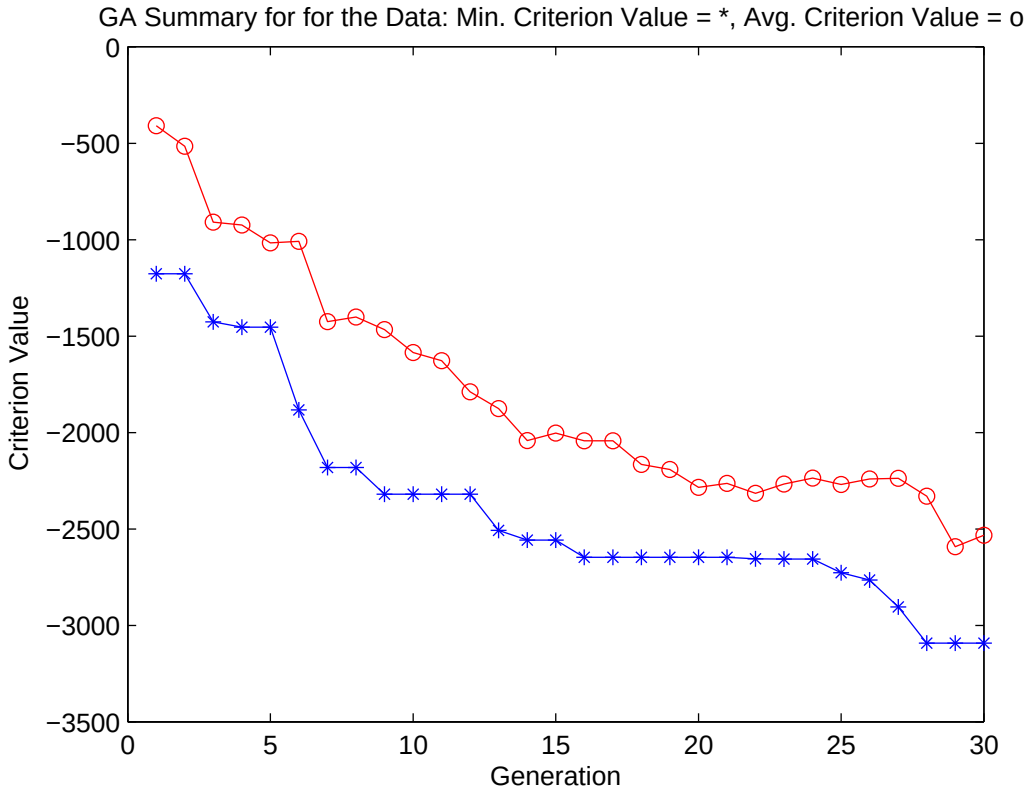

Figure 20: Second GA Run for the Test Data

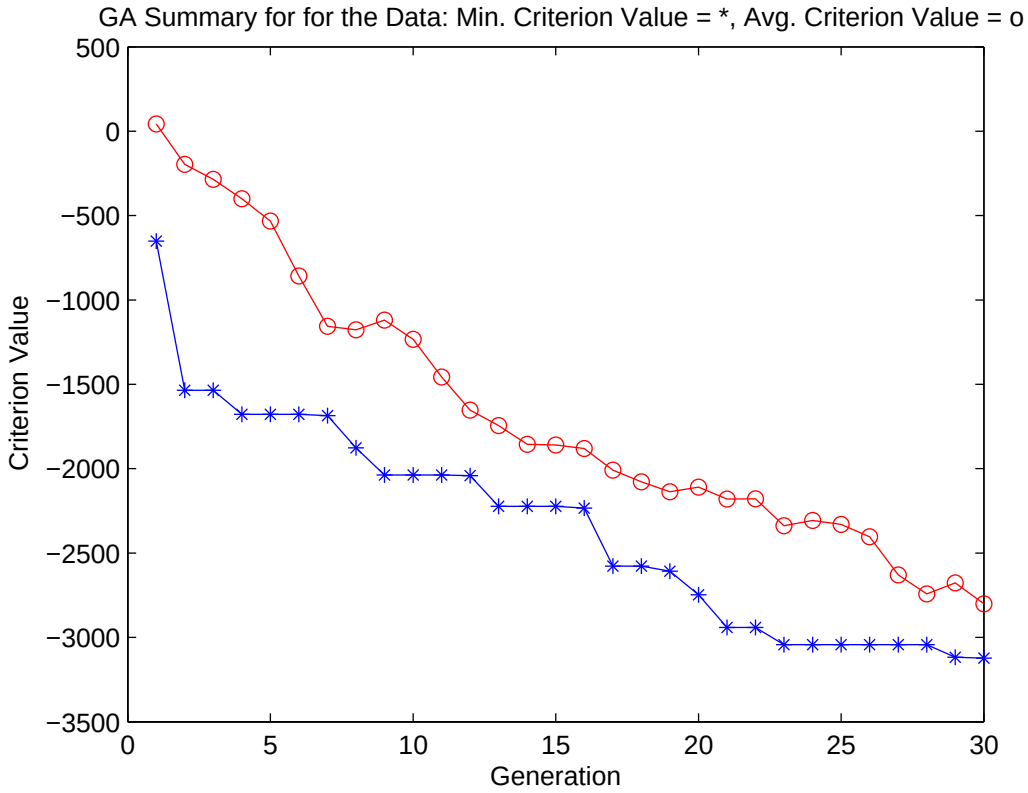

Figure 21: Third GA Run for the Test Data

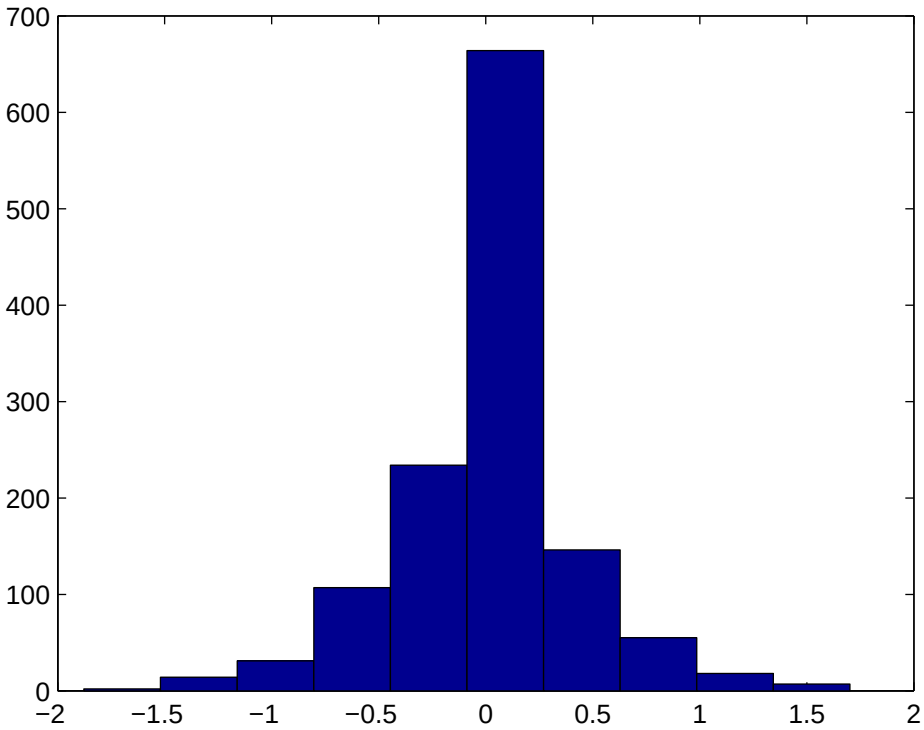

Figure 22: Distribution of Factor Score Residuals for the Test Data

VITA

I was born in Shangqiu, a small city in the southeast part of China. I spent my childhood and teenage years in Shangqiu with my parents and my brother. In 1996, I began to attend university in Xi'an, a major historic city in China. After graduation, I worked as an instructor of English in a small private college before I came to the US for my graduate study. I was admitted to the University of Tennessee's PhD program in Educational Psychology. My PhD advisor was Prof. Schuyler W. Huck. During my doctoral work, I also earned a MS degree in Statistics through the Intercollegiate Graduate Statistics Program chaired by Prof. Mary Sue Younger. And my MS in Statistics advisor was Prof. Hamparsum Bozodgan. My doctoral committee was co-chaired by Prof. Huck and Prof. Bozdogan and they were joined by Prof. Tricia McClam and Prof. Russell Zaretzki.