12-2008

Multivariate Mixed Data Mining with Gifi System using Genetic Algorithm and Information Complexity

Suman Katragadda
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Suman Katragadda entitled "Multivariate Mixed Data Mining with Gifi System using Genetic Algorithm and Information Complexity." 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 Business Administration.

Hamparsum Bozdogan, Major Professor

We have read this dissertation and recommend its acceptance:

Mary Leitnaker, Russell Zaretski, Mohammed Mohsin, Adam Petrie

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

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

To the Graduate Council:

I am submitting herewith a dissertation written by Suman Katragadda entitled "Multivariate Mixed Data Mining with Gifi System using Genetic Algorithm and Information Complexity " 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 Business Administration.

Hamparsum Bozdogan
Major Professor

We have read this dissertation
and recommend its acceptance:

Mary Leitnaker

Russell Zaretski

Mohammed Mohsin

Adam Petrie

Accepted for the Council:

Vice Provost and Dean of the
Graduate School

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

Multivariate Mixed Data Mining with Gifi System using Genetic Algorithm and Information Complexity

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Suman Katragadda

December 2008

Dedication

This dissertation is dedicated to my family - father, Nagabhushanam Katragadda, mother, Devaki Devi Katragadda, wife, Yamuna Katragadda, brother, Gopi Krishna Katragadda and brothers wife, Lakshmi Katragadda.

Acknowledgments

I would like to acknowledge, first, my Prof. Hamparsum Bozdogan for guiding my dissertation from the beginning to the very end. His thorough and efficient ongoing suggestions made this dissertation an excellent work for a student like me at the doctoral level. Without his sincere supervision of my research, this dissertation would not have been possible.

I would like to thank Dr. Mary Leitnaker, Dr. Russell Zaretski, Dr. Mohammed Mohsin and Dr. Adam Petrie for agreeing to be on my committee and also my special thanks to Dr. Mary Leitnaker for providing financial support for two consecutive years.

I would also like to acknowledge the University of Tennessee's Department of Statistics, Operations and Management Science for three and a half years of financial support. I would like to thank my parents, wife and relatives for always being there for me, giving moral support. I would also like to thank all those who have contributed directly or indirectly for the successful completion of this dissertation.

Abstract

Statistical analysis is very much dependent on the quality and type of a data set. There are three types of data - continuous, categorical and mixed. Of these three types, statistical modeling on a mixed data had been a challenging job for a long time. This is due to the fact that most of the traditional statistical techniques are defined either for purely continuous data or for purely categorical data but not mixed data. In reality, most of the data sets are neither continuous nor categorical in a pure sense but are in mixed form which makes the statistical analysis quite difficult. For instance, in the medical sector where classification of the data is very important, presence of many categorical and continuous predictors results in a poor model. In the insurance and finance sectors, lots of categorical and continuous data are collected on customers for targeted marketing, detection of suspicious insurance claims, actuarial modeling, risk analysis, modeling of financial derivatives, detection of profitable zones etc.

In this work, we bring together several relatively new developments in statistical model selection and data mining. In this work, we address two problems. The first problem is to determine the optimal number of mixtures from a multivariate Bernoulli distributed data using genetic algorithm and Bozdogan's information complexity, ICOMP. We show that the results of the maximum likelihood values are not just sufficient in determining the optimal number of mixtures. We also address the issue of high dimensional binary data using a genetic algorithm to determine the optimal predictors. Finally, we show the results of our algorithm on a simulated and two real data sets.

The second problem is to discovering interesting patterns from a complicated mixed data set. Since mixed data are a combination of continuous and categorical variables, we transform the non linear categorical variables to a linear scale by a mechanism called Gifi transformation, [Gifi, 1989]. Once the non linear variables are transformed to a linear scale (Euclidean space), we apply several classical multivariate techniques on the transformed continuous data to identify the unusual patterns. The advantage with this transformation is that it has a one-to-one mapping mechanism. Hence, the transformed set of continuous value(s) in the Gifi space can be remapped to a unique set of categorical value(s) in the original space. Once the data is transformed to the Gifi space, we implement various statistical techniques to identify interesting patterns. We also address the problem of high dimensional data using genetic algorithm for variable selection and Bozdogan's information complexity (ICOMP) as our fitness function.

We present details of our newly-developed Matlab toolbox, called *Gifi System*, that implements everything presented, and can readily be extended to add new functionality. Finally, results on both simulated and real world data sets are presented and discussed.

Keywords: *Gifi, homals, regression, multivariate logistic regression, fraud detection, medical diagnostics, supervised classification, unsupervised classification, variable selection, high dimensional data mining, stock market trading, detection of suspicious insurance claim estimates.*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statistical analysis is very much dependent on the quality and type of the data set. There are three types of data sets such as continuous, categorical and mixed. A continuous data set is one in which all the variables in it are in continuous form. A categorical data set is one in which all the variables in it are either ordinal (ordering of the categories exists) or nominal (no specific ordering of the categories exists). A particular form of categorical data set is a binary data set in which all the variables take values 0 and 1's. A mixed data set is one which contains some of the variables in continuous form and the rest of the variables in categorical form. In other words, a mixed data set is a combination of continuous and categorical data variables. Statistical analysis would have been easy if data set is purely continuous or purely categorical. In reality, most of the data sets are neither purely continuous nor purely categorical but are in mixed form which makes the statistical analysis quite difficult.

In this work, we address two problems. The first problem is about determining optimal number of clusters in a high dimensional binary data set. We can represent any data set in binary form by discretizing the continuous and the categorical variables (having more than two levels) by using suitable discretization procedures. We assume that the binary data set is Multivariate Binary distributed. We determine the optimal number of mixtures (clusters) using the information complexity, ICOMP ([Bozdogan, 1987], [Bozdogan, 1988], [Bozdogan, 90a], [Bozdogan, 90b], [Bozdogan, 2004]), as our selection criteria. We also address the problem of selecting the optimal number of variables from a high dimensional binary data set using the GA technique introduced by [Goldberg, 1989].

The second problem is about discovering some interesting patterns in a complicated mixed data set. Since a mixed data set is combination of continuous and categorical variables, we transform the non linear categorical variables to a linear scale by a mechanism called Gifi transformation, [Gifi, 1989]. Once the non linear variables are transformed to a linear scale (Euclidean space), we apply several classic multivariate techniques on the transformed continuous data set to identify the unusual patterns. Since the Gifi transformation has a one-to-one mapping of the nonlinear values to a linear value, the advantage of this transformation is that the final predicted results can be reverse mapped to the original scale from the transformed scale.

This dissertation is organized as follows. We briefly review the literature of the above two mentioned problems in chapter 2. Also in chapter 2, we review the concepts of Information Complexity and Genetic Algorithm. We also review the literature on categorical data coding and homogeneity analysis in this chapter. In chapter 3, we briefly discuss the concepts of Multivariate Bernoulli Distribution. We show that the maximum likelihood values are not just sufficient in determining the optimal number of mixtures in the Multivariate Binary Distributed data. We show that the optimal number of mixtures are selected by the maximum likelihood values in addition with the Information Complexity criteria, ICOMP. We also address the problem of selecting optimal number of variables in the model. We provide a solution using the genetic algorithm to select the optimal number of variables from a high dimensional binary data set.

In chapter 4, we describe the problems faced by statistician in a mixed data set and illustrate a procedure to handle such problems. In chapter 5, we illustrate several applications of the multivariate statistical methods in the Gifi space. For each application, we provide two algorithms - one with optimal scaling of the categorical variables in the Gifi space and the other with a linear combination of the categories of the categorical variables in the Gifi space. Numerical Results are reported in chapter 6.

Chapter 2

Literature Review

In this section, we review the literature on Multivariate Bernoulli distribution, Gifi system, Information Complexity, Genetic Algorithm, Homogeneity Analysis and Categorical Data Coding.

2.1 Multivariate Bernoulli Distributed Data

A binary variable is one that can take values 0 and 1 which indicates the absence and presence of that variable respectively. Let P be a population (multivariate binary data) consisting of n objects where each object is an observation on each of the J binary variables. Cluster Analysis, [Aderberg, 1973], is a technique of grouping these n objects from population P into one or more groups such that the objects within each group are similar and the objects between each group are quite dissimilar. Multivariate binary data arises in most of the disciplines such as chemistry, pharmacology, ecology, genetics, and social science, [Larsen and Liu, 2005].

A finite mixture model is one that comprises of two or more finite probability density functions ([Titterington et al., 1985], [McLachlan and Peel, 2000], [Lindsay, 1995]). Finite mixtures of multivariate Bernoulli distributions have been extensively used in diverse fields. In finite mixture modeling, most of the emphasis in the literature had been on Gaussian mixture models ([Dasgupta, 1999], [Dasgupta and Schulman, 2000], [Arora and Kannan, 2001]) and little attention had been on Bernoulli mixture models. One of the reasons might be that the class of finite mixtures of multivariate Bernoulli distributions are known to be nonidentifiable i.e., different values of the mixture parameters can correspond to exactly the same probability distribution. [Carreira-Perpinan and Renals, 2000] gave an empirical support to the fact that estimation of this class of mixtures can still produce meaningful results in practice, thus lessening the importance of the identifiability problem.

[Carreira-Perpinan, 2001] discussed the problem of finite mixtures of Bernoulli distributions where the selection of optimal number of mixtures is based on the minimum lack of fit criteria. We show that the minimum lack of fit criteria is not just sufficient in determining the optimal number of mixtures. We use the concepts of information complexity, [VanEmden, 1971], in selecting the best model. All information complexity criteria penalize a bad fitting model with negative twice the maximized log-likelihood, as an estimate of the Kullback-Liebler information ([Kullback, 1968], [Kullback and Leibler, 1951]). The difference then, is in the penalty for model complexity. We show that the information criteria, ICOMP, [Bozdogan, 1987], together with the lack of fit can determine the optimal number of mixtures in this case.

High dimensional data had been a problem by many researches in cluster analysis. It might be computationally expensive and convergence to the optimal parameter values might be time consuming. Moreover, not all predictors in the data might be needed for classification into the target number of mixtures. Selecting the optimal number of predictors from such a large number of predictors might be a challenging problem. We address the problem of high dimensional binary data by implementing the genetic algorithm ([Goldberg, 1989], [Holland, 1992], [Forrest, 1993], [Srinivas and Patnaik, 1994]). A genetic algorithm (GA) is a stochastic search algorithm which is based on concepts of biological evolution and natural selection that can be applied to solving problems where vast number of possible solutions exists.

2.2 Gifi System

Now, we review the literature on Gifi transformation, [Gifi, 1989]. Even though Alfred Gifi had written a book on this transformation in 1989, not much work has been done in this area till now. About a decade ago [Michailidis and de Leeuw, 1996] reviewed the concepts of Gifi transformation applied on a pure categorical data set. It was shown in detail the application of several classical multivariate techniques on the transformed scale to identify patterns in the categorical data set. From this we are motivated to apply the Gifi transformation on a mixed data set which is much more complicated than a pure categorical data set. In a mixed data setting, we apply the Gifi transformation on the qualitative variables leaving the continuous variables intact. After the transformation, the data set is no more of the mixed data type. It would be purely continuous in nature. Then, we apply the standard multivariate technique on this transformed continuous space to identify some useful patterns. Even in this problem, the best model and the optimal choice of variables to be included in the model are identified by ICOMP and the GA technique respectively.

2.3 Information Complexity

2.3.1 Introduction

The word information complexity involves notions such as connectivity patterns and the interactions of model components. In general statistical modeling and model evaluation problems, the concept of model complexity plays an important role ([Bozdogan, 2004]). Without considering the overall complexity of the model, its prediction and the goodness of fit of the model is difficult to assess. The art of selecting good statistical model lies in selecting a model that has minimum complexity from the vast pool of other possible models. In the sections that follow, we describe in detail some of the popular information criteria that are used in evaluating and selecting a good model from various other possible models. All information complexity criteria penalize a bad fitting model with negative twice the maximized log-likelihood, as an estimate of the *Kullback-Liebler Information*, [Kullback, 1968], [Kullback and Leibler, 1951].

2.3.2 AIC

AIC stands for Akaike's Information Criterion. AIC is the first information criterion introduced by Akaike in 1973 for model selection, [Akaike, 1973]. In general, AIC for model M_k is given by

$$AIC(M_k) = -2lnL(\hat{\theta}) + 2k, \tag{2.1}$$

where k is the number of independent parameters in the model M_k. The $-2lnL(\hat{\theta})$ is called the lack of fit component and $2k$ is the penalty component. The penalty component is a measure of complexity that compensates for the bias in the lack of fit when the maximum likelihood estimators are used ([Bozdogan, 1987]).

2.3.3 CAIC

CAIC stands for Consistent Akaike's Information Criterion. CAIC was developed by [Bozdogan, 1987] as an extension to AIC, [Akaike, 1973], to make it consistent without violating Akaike's underlying principles. In general, CAIC for model M_k is given by

$$CAIC(M_k) = -2lnL(\hat{\theta}) + k[ln(n) + 1], \tag{2.2}$$

where k is the number of independent parameters in the model and n is the number of observations in the data set. Another form of CAIC which takes in the inverse Fisher Information matrix (known as CAICF) is also introduced by [Bozdogan, 1987]. The form of CAICF for model M_k is given by

$$CAICF(M_k) = -2lnL(\hat{\theta}) + k[ln(n) + 2] + ln|\hat{F}|, \tag{2.3}$$

where $|\hat{F}|$ is the determinant of the estimated Fisher Information matrix, $\hat{F}$.

2.3.4 MDL \ SC

MDL stands for Minimum Description Length and SC stands for Schwarz criterion. MDL\SC penalize over-parameterized models more stringently than AIC. MDL\SC for model M_k is given by,

$$MDL\backslash SC(M_k) = -2lnL(\hat{\theta}) + k[ln(n)], \tag{2.4}$$

[Schwarz, 1978] suggested the criterion given in equation (2.4) assuming that the data is generated from an exponential family of distributions and using Bayes' procedure for the choice of a model. [Rissanen, 1978], [Rissanen, 1989] proposed a criterion based on information-theoretic shortest code length for the data together with the parameters of a model. His criterion is called MDL which is also given in equation 2.4. It is to be noted that MDL is identical to SC in form, but its derivation is quite different.

2.3.5 SBC

SBC stands for Schwarz Bayesian Criterion. SBC, [Schwarz, 1978], for model M_k is given by

$$SBC(M_k) = -2lnL(\hat{\theta}) + k[ln(n)], \tag{2.5}$$

2.3.6 ICOMP

ICOMP stands for Information Complexity. ICOMP ([Bozdogan, 1988], [Bozdogan, 90a], [Bozdogan, 90b]) was developed for model selection in general multivariate linear and nonlinear structural models. The formulation of ICOMP was motivated by AIC but it is based on the generalization of the information-based covariance complexity ([VanEmden, 1971]). For a general multivariate linear or nonlinear model defined by

$$StatisicalModel = Signal + Noise, \tag{2.6}$$

ICOMP is designed to estimation a loss function:

$$Loss = Lackoffit + LackofParsimony + Profusion \qquad of \qquad Complexity \tag{2.7}$$

in several ways using the additivity properties of information theory, [Bozdogan, 2004] and the developments of *Final Estimation Criterion (FEC)*, [Rissanen, 1976], for estimation and model identification problems, as well as AIC, [Akaike, 1973], and its analytical extensions ([Bozdogan, 1987]).

The development and construction of ICOMP is based on a generalization of the covariance complexity index, [VanEmden, 1971]. Unlike AIC, ICOMP penalizes the covariance complexity of the model instead of penalizing the free parameters directly. ICOMP for model M_k is given by

$$ICOMP(M_k) = -2lnL(\hat{\theta}) + 2C(\hat{\Sigma}_{M_k}) \tag{2.8}$$

where C is a real valued complexity measure and $\hat{Cov}(\hat{\theta}) = \hat{\Sigma}_{M_k}$ represents the estimated covariance matrix of the parameters of the model M_k.

There are several forms of ICOMP defined in the literature. The first form of ICOMP is given by

$$ICOMP = -2lnL(\hat{\theta}) + 2C_1(\hat{Cov}(\hat{\theta})), \tag{2.9}$$

where $C_1(\hat{Cov}(\hat{\theta}))$ is the maximal information theoretic measure of complexity of a covariance matrix $\hat{Cov}(\hat{\theta})$. $C_1(\hat{Cov}(\hat{\theta}))$ is given by

$$C_1(\hat{Cov}(\hat{\theta})) = \frac{p}{2}log[\frac{tr(\hat{Cov}(\hat{\theta}))}{p}] - \frac{1}{2}log|\hat{Cov}(\hat{\theta})|, \tag{2.10}$$

where p is rank of the covariance matrix, $\hat{Cov}(\hat{\theta})$.

A variant of the first form of ICOMP using the second order equivalent measure of complexity, C_{1F} to the original C_1 is given by

$$ICOMP = -2lnL(\hat{\theta}) + 2C_{1F}(\hat{Cov}(\hat{\theta})), \tag{2.11}$$

where C_{1F} is given by

$$\begin{aligned} C_{1F}(\hat{Cov}(\hat{\theta})) &= \frac{s}{4}\frac{(1/s)tr(\hat{Cov}(\hat{\theta})\hat{Cov}(\hat{\theta})') - (\frac{tr(\hat{Cov}(\hat{\theta}))}{s})^2}{(\frac{tr(\hat{Cov}(\theta))}{s})^2} \\ &= \frac{1}{4\bar{\lambda}^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda})^2 \cong C_1(\hat{Cov}(\hat{\theta})) \end{aligned} \tag{2.12}$$

where λ_j's are the eigenvalues of $\hat{Cov}(\hat{\theta})$ for $j = 1, 2, ..., s$ and $\bar{\lambda}$ is the arithmetic mean of the eigenvalues. $C_{1F}(.)$ is scale-invariant and $C_{1F}(.) \geq 0$. It measures the relative variation in the eigenvalues.

The second form of ICOMP which uses the complexity of the estimated inverse-Fisher information matrix (IFIM), $\hat{F}^{-1}$, is given by

$$ICOMP(IFIM) = -2lnL(\hat{\theta}) + 2C_1(\hat{F}^{-1}) \tag{2.13}$$

A variant of the second form of ICOMP using the second order equivalent measure of complexity, C_{1F} to the original C_1 is given by

$$ICOMP = -2lnL(\hat{\theta}) + 2C_{1F}(\hat{F}^{-1}). \tag{2.14}$$

There is another form of ICOMP, known as $ICOMP_{PEU}$, which is an approximation of the posterior expected utility (PEU). It is a useful form of ICOMP in modeling situations characterized by over parameterization. It clearly enforces a stricter penalty than the usual ICOMP. It is given by

$$ICOMP_{PEU\_LN} = -2lnL(\hat{\theta}) + k + 2C_1(\hat{F}^{-1}). \tag{2.15}$$

The consistent $ICOMP_{PEU}$ is given by

$$ICOMP_{PEU\_LN} = -2lnL(\hat{\theta}) + k + ln(n)C_1(\hat{F}^{-1}). \tag{2.16}$$

ICOMP is also defined under model misspecification. When a model is misspecified, ICOMP under misspecification is given by

$$ICOMP(IFIM)_{Misspec} = -2ln(L(\hat{\theta})) + 2C_1(\hat{Cov}(\hat{\theta})_{Misspec}), \tag{2.17}$$

where $\hat{Cov}(\hat{\theta})_{Misspec} = \hat{F}^{-1}\hat{R}\hat{F}^{-1}$. The matrix $\hat{R}$ is the estimated outer product form of the inverse Fisher Information matrix (IFIM). $\hat{Cov}(\hat{\theta})_{Misspec}$ is a consistent estimator of $\hat{Cov}(\hat{\theta})$. This is often called the sandwich covariance or robust covariance estimator, since it is a correct variance regardless whether the assumed model is correct or not. ICOMP under misspecification enforces even higher penalty term than the other versions of ICOMP.

2.4 Genetic Algorithm

2.4.1 Introduction

Genetic Algorithms (GA) are adaptive heuristic search algorithms premised on the evolutionary ideas of natural selection and genetic. They are a particular class of evolutionary algorithms (also known as evolutionary computation) that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover (also called recombination). They are used in computing to find exact or approximate solutions to optimization and search problems. Genetic Algorithms are categorized as global search heuristics. A detailed review on GA and its importance is given in [Marczyk, 2004] and [Mangano, 1996] respectively.

2.4.2 Methodology

A GA is a stochastic search algorithm which is based on concepts of biological evolution and natural selection that can be applied to solving problems where vast number of possible solutions exists. Unlike conventional optimization techniques, the GA requires no calculation of the gradient of the objective function and is not restricted to local optima ([Goldberg, 1989]).

A GA treats information as a series of codes on a binary string, where each string represents a different solution to a given problem. These strings are analogous models to the genetic information coded by genes on chromosome. A string can be evaluated according to some fitness value, for its particular ability to solve the problem. On the basis of the fitness values, strings are either retained or removed from the analysis after each run so that, after many runs, the best solution have been identified. One important difficulty with any GA is in choosing an appropriate fitness function as the basis for evaluating each solution.

2.4.3 Crossover

Mating is performed as a crossover process ([Bozdogan, 2004]). A model chosen for crossover is controlled by the crossover probability or the crossover rate. The crossover probability is often determined by the investigator. A crossover probability of zero simply means that the members of the mating pool are carried over into the next generation and no off springs are produced. A crossover probability of one indicates that the mating (crossover) always occurs between any two parent models chosen from the mating pool; thus the next generation will consist only of off spring models (not of any models from the previous generation).

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

In this work, we discuss three different crossover methods that can be performed.

Single Point Crossover

One crossover point is selected; binary string from beginning of the chromosome to the crossover point is copied from one parent, the rest is copied from the second parent. For example,

Parent A:	1	1	0	0	|	1	0	1	1
Parent B:	1	1	0	1	|	1	1	1	1
	–	–	–	–	–	–	–	–	
Offspring1:	1	1	0	0	|	1	1	1	1
Offspring2:	1	1	0	1	|	1	0	1	1
	–	–	–	–	–	–	–	–	–

Two Point Crossover

Two crossover points are selected; binary string from the beginning of the chromosome to the first crossover point is copied from one parent, the part from the first to the second crossover point is copied from the second parent and the rest is copied from the first parent. For example,

Parent A:	1	1	0	|	0	1	0	|	1	1
Parent B:	1	1	0	|	1	1	1	|	1	1
	–	–	–	–	–	–	–	–	–	–
Offspring1:	1	1	0	|	1	1	1	|	1	1
Offspring2:	1	1	0	|	0	1	0	|	1	1
	–	–	–	–	–	–	–	–	–	–

Uniform Crossover

In this, the bits are randomly copied from the first or from the second parent. For example,

Parent A:	1	1	0	0	1	0	1	1
Parent B:	1	1	0	1	1	1	1	1
	–	–	–	–	–	–	–	–
Offspring1:	1	1	0	1	1	1	1	1
Offspring2:	1	1	0	0	1	0	1	1
	–	–	–	–	–	–	–	–

In our algorithm, the user has the option of choosing any one of the three crossovers.

2.4.4 Mutation

Mutation of models is used in GA as another means of creating new combinations of variables so that the searching process can jump to another area of the fitness function landscape instead of searching in a limited area. By mutation, a randomly selected locus can change from 0 to 1 or from 1 to 0. Thus, a randomly selected predictor variable is either added to or removed from the model.

2.4.5 GA process

A GA procedure contains the following steps.

Step 1:

Each subset of the data set is coded as a series of 1's and 0's (binary string). For example, if we have 10 predictors in the data set with labels A to J. The binary string 1001000101 represent the dataset which contains predictors A, D, H and J.

Step 2:

Randomly generate the initial population. An initial population of size N contains N binary strings where each string represents a subset of the original data set.

Step 3:

Evaluate each member in the population by an appropriate fitness function. Any model selection criteria such as AIC, SBC, ICOMP, $ICOMP_{PEU}$, $ICOMP_{PEU_{LN}}$, CAIC, $ICOMP_{Misspec}$ can be used as a fitness function. In this work, we use the variants of ICOMP as the fitness function.

Step 4:

Create new population by performing mating of parent models to reproduce offspring models. Crossover and Mutation are performed on the parent models to reproduce offspring models.

Step 5:

Repeat steps 1 to 4 up to the maximum number of iterations desired by the investigator (stopping criteria).

2.4.6 Pseudo code

We briefly describe the algorithm for selecting the new population.

Regular procedure

1. Generate a random population of N random models.
2. Evaluate each model in the population by using the fitness function.
3. Select two parent models from the population (Selection).
4. Perform Crossover operation with a crossover probability on the two parent models to produce an offspring.
5. Perform Mutation operation with a mutation probability on the offspring and place it in the new population.
6. Perform steps 3, 4 and 5 N times so that the new population has N new models in it.
7. If the stopping criteria is met, return the best solution from the current population
8. Go to Step 2

Slight modification of the regular procedure

1. Generate a random population of N random models.

2. Evaluate each model in the population by using the fitness function.

3. Sort the models in the population in the increasing order of the fitness function; the model with the minimum fitness function as the first element in the population.

4. Since the population has N models, we choose the first $N/2$ models for the crossover operation.

5. Crossover operation is performed on each of the models in the population from 2 to $N/2$ with the first model.

6. The new population always contains the first model in the population; the model having the minimum fitness function.

7. The $N-2$ offspring's produced in step 4 go in to the new population. At this point we have $N-1$ models in the new population.

8. To generate a new population of size N, we perform crossover with the first model in the old population and the $N/2+1$ model in the old population. Since there will be two offspring's produced from this crossover operation, we randomly select one offspring and place it in the new population. Hence the new population contains N models.

9. Perform Mutation operation with a mutation probability on the models in the new population.

10. If the stopping criteria is met, return the best solution (which is the first model) from the current population

11. Go to Step 2

2.4.7 Advantages of GA

There are many advantages of using GA as a search algorithm. Some of them are briefly listed below.

- GA can be used in parallel processing.

- GA results in giving the optimal or near optimal solutions where the solution space is vast.

- GA performs well on complex fitness functions. Complex fitness functions are those that are discontinuous, noisy, changes over time, or have many local optima ([Marczyk, 2004]).

- They can be used in searching for optimal solutions (or near optimal solutions) where simultaneous computation on multi parameters is needed ([Forrest, 1993]).

2.4.8 Disadvantages of GA

Some of the main disadvantages of GA are briefly listed below.

- Computational intensive

- Resource intensive

- Choice of a good fitness function is appropriate

- Parameter inputs regarding population size, mutation probability, and crossover probability must be considered with care.

2.4.9 Applications of GA

GAs are used in variety of disciplines. Some of them are listed below.

- Finance

- Economics

- Game Programming

- Robotics

- Mathematics

- Pattern Recognition and Data Mining

- Genome Science

- Chemistry
- Astrophysics
- Astronomy
- Resource allocation
- Scheduling
- Routing
- Music Theory

2.4.10 Example: Body Fat Data

We illustrate the results of the GA on a real data set, Body Fat. The body fat data consists of 15 variables and 252 observations. The 15 variables are briefly listed below.

- X1: Density determined from underwater weighing
- Y: Percent body fat
- X2: Age (years)
- X3: Weight (lbs)
- X4: Height (inches)
- X5: Neck circumference (cm)
- X6: Chest circumference (cm)
- X7: Abdomen 2 circumference (cm)
- X8: Hip circumference (cm)
- X9: Thigh circumference (cm)
- X10: Knee circumference (cm)
- X11: Ankle circumference (cm)

- X12: Biceps (extended) circumference (cm)
- X13: Forearm circumference (cm)
- X14: Wrist circumference (cm)

The overall goal is to estimate the percent body fat given the other variables as predictors in the model. We fit a multiple regression model with Y as dependent variable and X1-X14 as independent variables. We include the intercept term in the model. We run the genetic algorithm with the following input parameters.

Number of iterations: 100
Population Size: 20
Crossover: 0.75
Mutation: 0.10
Crossover: Uniform
Fitness function: $ICOMP_{C1}$

The following predictors are selected as the best predictors.

Model: Intercept, X1, X6

The $ICOMP_{C1}$ score for this model is 852.8459. The r-square value for this model is 97.72%. The parameter coefficients for this model are given by

$$\beta = \begin{bmatrix} 455.0651 \\ -418.0924 \\ 0.0536 \end{bmatrix}$$

The standard error of the parameter estimates are given by

$$S_{\beta} = \begin{bmatrix} 6.9898 \\ 5.7186 \\ 0.0129 \end{bmatrix}$$

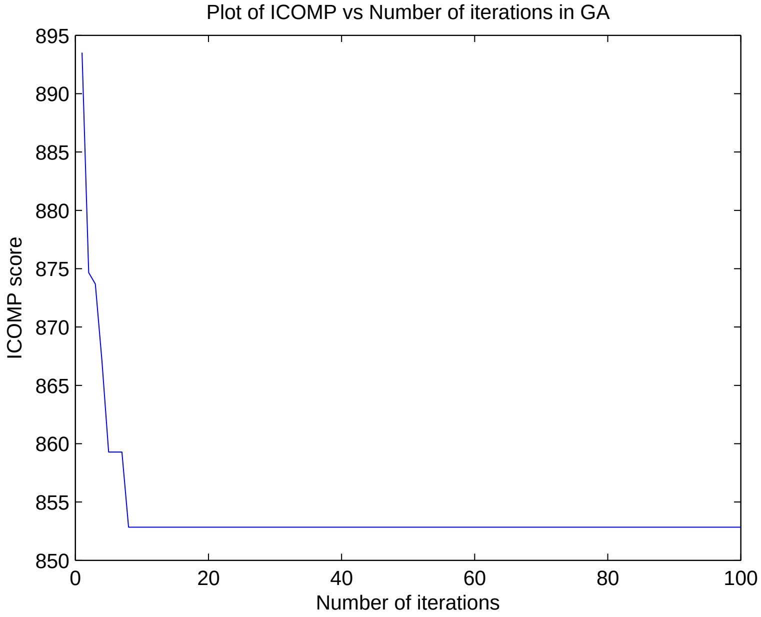


Figure 2.1: Body Fat Data: Plot of ICOMP vs Number of iterations in GA.

The regression sum of squares is computed as 17178 and the total sum of squares is computed as 17579. The F ratio for this model is computed to be 5337.1. The best (minimum) value of $ICOMP_{C1}$ at the end of each iteration is shown in Figure 2.1. When we run this data using standard regression model in NCSS, the following parameters are selected.

Model: Intercept, X1, X2

This model produced an R-square of 97.61% and the $ICOMP_{C1}$ score for this model is 864.1535.

2.5 Coding of Categorical Data

2.5.1 Introduction

Categorical variable (also known as qualitative variable) is a type of data which may be divided into categories or groups. For instance, the variable gender has only two categories namely male and female. Hence, it is termed as a categorical variable. Categorical variables are discrete in nature. There are two types of categorical variables namely ordinal and nominal. An ordinal variable, [Tamhane and Dunlop, 2003], is a type of categorical variable where the categories of that variable can be ordered or ranked (e.g., Disagree, Neutral, Agree). A nominal variable is a type of categorical variable where the categories simply represent distinct labels (e.g., Red, Green, Black).

2.5.2 Data Representation

Let us assume that there are finite number of m categorical variables h_j $(j = 1, 2, ..., m)$. Also assume that each variable h_j has k_j distinct categories. Suppose that a finite set of n objects (or individuals) are collected on these m categorical variables. We represent the data matrix H as an $n \times m$ matrix with elements h_{ij} giving the category of variable h_j for object i.

Bipartite Graph

Given such a data matrix H, one can represent all the available information by a bipartite graph, [Michailidis and de Leeuw, 1996], where the first set of n vertices corresponds to the objects and the second set of $\sum_{j=1}^{m} k_j$ vertices to the categories of the m variables. Each object is connected to the categories of the variables it belongs to. Hence the set of $n \sum_{j=1}^{m} k_j$ provides information about which categories an object belongs to, or alternatively which objects belong to a specific category. The n vertices corresponding to the objects all have degree m, while the $\sum_{j=1}^{m} k_j$ vertices corresponding to the categories have varying degrees, equal to the number of objects in the categories. For instance, if data on two categorical variables are collected on 5 objects where the number of categories for the first variable is two and the number of categories for the second variable is 3. A bipartite graph for this data is shown in Figure 2.2.

A bipartite graph would be of minimum use if n and m are large since the graph might have too much ink and might be difficult to identify interesting patterns.

Binary coding

Another way to represent the data matrix H is to represent each variable h_j as a binary indicator matrix which is described in detail in the next section.

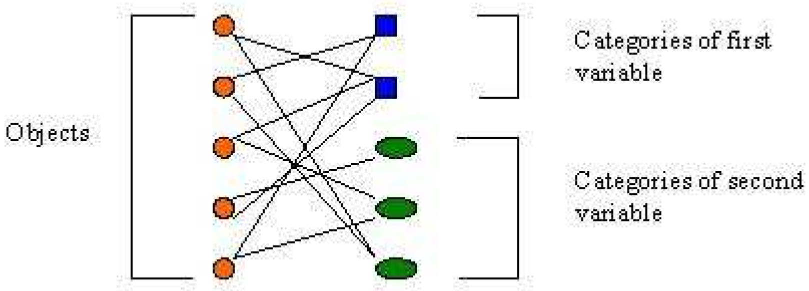


Figure 2.2: Example of a Bipartite Graph.

Table 2.1: Data matrix: H

p	x	a
p	x	b
p	x	a
q	x	a
q	x	b
p	y	b
q	y	a
p	x	b
q	x	a
p	x	a

2.5.3 Indicator Matrix

An $n \times k_j$ binary matrix G_j for each variable h_j is defined as $G_j(i,t) = 1$, $i = 1, ..., N$, $t = 1, ..., k_j$ if object i belongs to category t, and $G_j(i,t) = 0$ if it belongs to some other category. G_j is called the *indicator matrix* of h_j. The matrix $G = (G_1, ..., G_j, ..., G_m)$ of dimension $n \times \sum_{j=1}^{m} k_j$ is a collection of such matrices and is also called an indicator matrix. Now, we illustrate an example of an indicator matrix. Consider a data matrix H, with 10 observations (or objects) and 3 categorical variables, given in Table 2.1. Each of the 3 categorical variables has two categories. The profile frequency of the data matrix H is given in Table 2.2 and the reduced profile frequency of H is given in Table 2.3. The indicator matrix G is given in Table 2.4.

Table 2.2: Profile Frequency of H

p	x	a	3
p	x	b	2
p	y	a	0
p	y	b	1
q	x	a	2
q	x	b	1
q	y	a	1
q	y	b	0

Table 2.3: Reduced Profile Frequency of H

p	x	a	3
p	x	b	2
p	y	b	1
q	x	a	2
q	x	b	1
q	y	a	1

Table 2.4: Indicator Matrix G for the data matrix H

p	q	x	y	a	b
1	0	1	0	1	0
1	0	1	0	0	1
1	0	1	0	1	0
0	1	1	0	1	0
0	1	1	0	0	1
1	0	0	1	0	1
0	1	0	1	1	0
1	0	1	0	0	1
0	1	1	0	1	0
1	0	1	0	1	0

Complete Indicator Matrix

The indicator matrix G_j is said to be complete if each row of G_j has only one element equal to unity and zeros elsewhere, so that row sums of G_j are equal to unity ([Gifi, 1989]). In vector form, we can represent this as $G_j u = u$ where u is a vector of unit elements. If all G_j are complete, their combined matrix G is also said to be complete. In vector form, we can write $Gu = mu$ since the rows of G add up to m.

Properties of a Complete Indicator Matrix

The properties of a complete indicator matrix are briefly described below.

1. Let d_j be the vector of column totals of G_j. Its kth element corresponds to the kth category of h_j. The sum of the elements in d_j must equal n. Mathematically, it can be written as $u^{'}d_j = n$ where u is a vector of 1's.

2. Since an object corresponds to only one category of the variable, the columns of the matrix G_j are orthogonal.

3. Let $D_j = G_j^{'}G_j$, be a diagonal matrix where the kth diagonal element equals the kth element in d_j. We define M_j as the diagonal matrix of row totals of G_j. For a complete indicator matrix, $M_j = Im$ where I is an identity matrix. We define $M_* = \sum M_j$ of h_j.

4. Let $C_{jl} = G_j^{'}G_l$, be a two dimensional cross tabulation of variables h_j and h_l. Its elements correspond to the frequency of objects characterized by a particular combination of one category in h_j and one in h_l. We define C as a combination of all C_{jl}'s. The jth diagonal sub-matrices in C corresponds to the diagonal matrix, D_j for variable h_j.

5. We define D as the partitioned matrix of C, in the sense that elements of D and C are identical in the diagonal sub-matrices $C_{jj} = D_j$, where D has zero elements in its off diagonal sub-matrices. D is a matrix of univariate marginals. The matrices C and D for the data matrix H are given in Tables 2.5 and 2.6 respectively.

Table 2.5: Matric C for the data matrix H

	p	q	x	y	a	b
p	6	0	5	1	3	3
q	0	4	3	1	3	1
x	5	3	8	0	5	3
y	1	1	0	2	1	1
a	3	3	5	1	6	0
b	3	1	3	1	0	4

Table 2.6: Matric D for the data matrix H

	p	q	x	y	a	b
p	6	0	0	0	0	0
q	0	4	0	0	0	0
x	0	0	8	0	0	0
y	0	0	0	2	0	0
a	0	0	0	0	6	0
b	0	0	0	0	0	4

Incomplete Indicator Matrix

An indicator matrix G_j is *incomplete* if it has rows with only zero elements. An incomplete indicator matrix can be quantified using the same principles outlined for the complete indicator matrix. Therefore,

$$x \propto M_*^{-1} G y$$

$$y_j \propto D_j^{-1} G_j^{'} x$$

Since the object scores will become more similar to the extent that the two objects have more categories in common, a solution based on the above requirements will be different from a solution based on the complete indicator matrix ([Gifi, 1989]).

Reversed Indicator Matrix

The reversed indicator matrix is derived from the transpose of the original indicator matrix G. We illustrate this with an example. Consider a data matrix H_1, with 5 objects and 2 categorical variables where each categorical variable has 3 levels, given in Table 2.7. The transposed data matrix of H_1 is given in Table 2.8. The reversed indicator matrix for H_1 is given in Table 2.9.

Table 2.7: Data Matrix, H_1

I	II
p	x
p	y
q	y
r	y
r	z

Table 2.8: Transposed Data Matrix, H_1

	1	2	3	4	5
I	p	p	q	r	r
II	x	y	y	y	z

2.5.4 Quantification

Quantification of a categorical variable h_j is a process of converting its categorical value to a continuous scale so that the classical techniques of multivariate analysis (MVA) can be applied. Quantification of categories of variable h_j implies that these k_j categories are mapped as the k_j numerical values of a vector y_j. Let the quantified variable, $q_j = G_j y_j$ be a single vector which gives a numerical result for each object with respect to h_j.

Let us define $x = m^{-1} \sum q_j$, the mean vector of all q_j's. x contains the quantification of the objects or in other words, the induced score of objects. We define the category quantification of a category as the average of the scores of those objects that are mapped into that category. Mathematically, we write it as $y_j = D_j^{-1} G_j^{'} x$. The vector x would be of size $n \times 1$ and the vector y_j is of size $k_j \times 1$.

2.5.5 Missing Data

The presence of missing data has been a recurring problem in multivariate data analysis. There might be many reasons for the presence of missing data. One such reason might be that a subject left a blank on his/her response sheet. Many ways of handling missing data have been proposed. One such proposal would be to insert a random value selected from

Table 2.9: Reversed Indicator Matrix for H_1

	1p	1x	2p	2y	3q	3y	4r	4y	5r	5z
I	1	0	1	0	1	0	1	0	1	0
II	0	1	0	1	0	1	0	1	0	1

Table 2.10: Data Matrix, H_2

p	x
p	z
q	y
p	⋆
r	⋆
⋆	x

Table 2.11: Indicator matrix with missing data

p	q	r	x	y	z
1	0	0	1	0	0
1	0	0	0	0	1
0	1	0	0	1	0
1	0	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0

the range of possible values. In this work, we shall distinguish the missing data with the following three options.

Using the indicator matrix, we can represent the missing data in three possible ways.

1. the indicator matrix is left incomplete.
2. the indicator matrix is completed with a single additional column for each variable with missing data.
3. the indicator matrix is completed by adding to G_j as many additional columns as there are missing data for the jth variable.

We illustrate the above three cases with an example. Consider a data matrix H_2 in Table 2.10. It has 6 objects on 2 categorical variables. Some of the objects have missing values. The indicator matrix for case 1 is shown in Table 2.11, for case 2 in Table 2.12 and for case 3 in Table 2.13 respectively.

2.6 Homogeneity Analysis

2.6.1 Introduction

The word homogeneity means the quality of being similar or comparable in nature. We say a data matrix H is homogeneous if and only if all the variables in H are similar. That is,

Table 2.12: Indicator matrix with missing data, single category

p	q	r	⋆	x	y	z	⋆
1	0	0	0	1	0	0	0
1	0	0	0	0	0	1	0
0	1	0	0	0	1	0	0
1	0	0	0	0	0	0	1
0	0	1	0	0	0	0	1
0	0	0	1	1	0	0	0

Table 2.13: Indicator matrix with missing data, multiple category

p	q	r	⋆	x	y	z	⋆	⋆
1	0	0	0	1	0	0	0	0
1	0	0	0	0	0	1	0	0
0	1	0	0	0	1	0	0	0
1	0	0	0	0	0	0	1	0
0	0	1	0	0	0	0	0	1
0	0	0	1	1	0	0	0	0

all the variables measure the same thing. In this sense, if we plot each observation in H as profiles, each profile would be a straight horizontal line. If the idea of measuring the same thing were imperfectly true (variables measure the same thing, but with random error), rows of H may have elements that vary somewhat (more to the extent that measurement error increases). A graph of profiles would then show zigzag curves at different levels. Replacing such profiles by a straight line then implies some loss of information. Variables are homogeneous if the loss is relatively small.

According to [Gifi, 1989], the term homogeneity analysis can be used in a strict sense and a broad sense. In a strict sense, it is a technique for the analysis of purely categorical data, with a particular loss function that defines it and with a particular method for finding an optimal solution. In a broad sense, it refers to a class of criteria for analyzing multivariate data in general, sharing the characteristic aim of optimizing the homogeneity of variables under various forms of manipulation and simplification.

According to [Michailidis and de Leeuw, 1996], the basic premise of homogeneity analysis was that complicated multivariate data can be made more accessible by displaying their main regularities and patterns in plots. The technique scales the n objects (map them into a low dimensional Euclidean space) in such a way that objects with similar profiles were close together, while objects with different profiles were relatively apart. In one way, this transformation optimally scales the categorical values to their corresponding continuous scores. Several multivariate techniques on nonlinear data are reported in [Breiman and Friedman, 1985], [Hastie et al., 1994], [Meulman and der Kooij, 2000], [Buuren and Heiser, 1989], [der Kooji and Meulman, 1997], [SPSS, 1999], [SPSS, 2004], [SPSS, 2006], [Young, 1981], [Young et al., 1976], [Young et al., 1978], [Gower and Blasius, 2005], [Groenen et al., 1998], [Guttman, 1941], [Heiser and Meulman, 1994], [Kruskal, 1964], [Meulman, 1982], [Meulman, 1992], [Meulman, 1993], [Meulman, 1996], [Meulman, 1998], [Meulman, 2003], [Meulman et al., 2002], [Meulman et al., 2004], [Nishisato, 1980], and [Nishisato, 1994].

In the next section, we explain in detail the working of HOMALS (Homogeneity Analysis by means of Alternating Least Squares). Most of the review of HOMALS is from [Gifi, 1989] and [Michailidis and de Leeuw, 1996].

2.6.2 HOMALS

Categorical PCA (HOMALS) is a particular form of nonlinear PCA that is based on a categorical coding of variables in indicator matrices. As described in the Categorical Data Coding chapter, G_j is an indicator matrix for variable j. The quantification of objects and of categories for a set of complete indicator matrices $\{G_1, ...G_j, ...G_m\}$ should satisfy the following proportionalities.

$$x \propto m^{-1} \sum_j G_j y_j \tag{2.18}$$

$$y_j \propto D_j^{-1} G_j^{'} x \tag{2.19}$$

In the equations (2.18) and (2.19), x is the vector of object scores and y_j is the vector of the quantifications of the categories of variable j.

Let X be the $n \times p$ matrix (usually $p \leq m$) containing the object scores and Y_j be the $k_j \times p$ matrix containing the category quantification of variable j. Since the quantification process incurs some loss of information, a typical loss function is given as,

$$\begin{aligned}\sigma(X;Y_1,...,Y_m) &= m^{-1}\sum_{j=1}^{m} SSQ(X-G_jY_j) \\ &= m^{-1}trace[(X-G_jY_j)'(X-G_jY_j)], \qquad (2.20)\end{aligned}$$

where SSQ(H) denotes the sum of squares of the elements of the matrix H. The loss function (2.20) is at the heart of the Gifi System ([Gifi, 1989]). We want to minimize the above loss function simultaneously over X and Y_j's. The entire system is mainly about different versions of the above minimization problem. By imposing various restrictions on the category quantifications Y_j and in some cases coding of the data, different types of analysis can be derived.

In the process of minimizing the loss function in (2.20), we impose two constraints in order to avoid the trivial solution corresponding to $X = 0$, and $Y_j = 0$ for every j. The two constraints are

$$\begin{aligned}X^{'}X &= nI_p \qquad (2.21) \\ u^{'}X &= 0, \qquad (2.22)\end{aligned}$$

where u is a vector of ones with dimension $p\times 1$. The constraint in (2.21) standardizes the squared length of the object scores (to be equal to n), and in two or higher dimensions also requires the columns of X to be in addition orthogonal. The constraint in (2.22) basically requires the graph plot to be centered around the origin.

We minimize the above loss function simultaneously over X and Y_j's by employing an *Alternating Least Squares* (ALS) algorithm. We start the process with a uniformly random choice of X ($X \neq 0$), with a mean zero, normalize it to the sum of squares n (rather than 1, so that the scores have variance 1). We compute a first set of category quantification Y_j by

$$\hat{Y_j} = D_j^{-1}G_j^{'}X \qquad (2.23)$$

where $D_j = G_j^{'}G_j$ is the $k_j \times k_j$ diagonal matrix containing the univariate marginals of variable j.

In the second step of the algorithm, the loss function in (2.20) is minimized with respect to X for fixed Y_j's. It is given by

$$\hat{X} = m^{-1} \sum_{j=1}^{m} G_j Y j. \tag{2.24}$$

In the third step of the algorithm the object scores X are column centered by setting $B = \hat{X} - u(u^{'}X/n)$, and then orthonormalized by the modified Gram-Schmidt procedure, [Trefethen and Bau, 1997], $X = \sqrt{n}GRAM(B)$, so that both the normalization constraints in (2.21) and (2.22) are satisfied. The usual normalization condition used in ALS is given by

$$X = X(X^{'}X)^{-1/2}. \tag{2.25}$$

The problem with the usual normalization condition in (2.25) might arise when p is large. When p is large this method could become quite expensive from a computational point of view. It can be replaced with the cheaper Gram-Schmidt method. The Gram-Schmidt method starts with unit normalizing the first column of X, then projects the second column of X onto the space orthogonal to the first column, replaces the second column by the unit normalized antiprojection, next projects the third column of X onto the space orthogonal to the new second column, and so on. This process can be summarized by stating that X is decomposed as $X = UT$, with $U^{'}U = I$ and T an upper triangular matrix. The matrix U is scaled by the $\sqrt{n}$ and the resulting matrix is taken as the new X.

The ALS algorithm cycles through these three steps until the convergence criterion is met. The first step in (2.23) expresses the first centroid principle (a category quantification is in the centroid of the object scores they belong to it), while the second step in (2.24) shows that an object score is the average of the quantifications of the categories it belongs to. Hence, this solution accomplishes the goal of producing a graph plot with objects close to the categories they fall in and categories close to the objects belonging in them.

Once the ALS algorithm has converged, by using the fact that $\hat{Y}_j' D_j \hat{Y}_j = \hat{Y}_j' D_j (D_j^{-1} G_j' \hat{X})$ $= \hat{Y}_j' G_j' \hat{X}$, we can write the loss function in (2.20) as

$$\begin{aligned} m^{-1} \sum_{j=1}^{m} tr[(\hat{X} - G_j \hat{Y}_j)^{'} (\hat{X} - G_j \hat{Y}_j)] &= m^{-1} \sum_{j=1}^{m} tr(\hat{X}' \hat{X} - \hat{Y}_j' D_j \hat{Y}_j) \\ &= m^{-1} \sum_{j=1}^{m} tr(nI_p - \hat{Y}_j' D_j \hat{Y}_j) \\ &= np - m^{-1} \sum_{j=1}^{m} tr(\hat{Y}_j' D_j \hat{Y}_j). \qquad (2.26) \end{aligned}$$

The sum of the diagonal elements of the matrices $\hat{Y}_j' D_j \hat{Y}_j$ is called the fit of the solution.

2.6.3 Discrimination Measures: Contribution of Variables

The discrimination measures in HOMALS are given for each variable in each dimension. The discrimination measure for the jth variable in sth dimension is given by

$$\eta_{js}^2 = y_{(j)s}^{'} D_j y_{(j)s}/n, \qquad (2.27)$$

where $y_{(j)s}$ is the quantification for h_j in the sth dimension of the solution.

The discrimination measures give the average squared distance (weighted by the marginal frequencies) of the category quantifications to the origin of the p-dimensional space. The discrimination measures add up across variables to $y_s^{'} D y_s/n = \psi_s^2$, so that the reported eigenvalue ψ_s^2/m is the average of the discrimination measures in the sth dimension. When a variable does not contribute to the sth dimension of the solution, the discrimination measure is zero (its category quantifications coincide with the origin). It can be shown that the discrimination measures are equal to the squared correlation between an optimally quantified variable $G_j Y_j(\hat{.}, s)$ in dimension s, and the corresponding column of object scores

$X(\hat{.,s})$. Hence, the loss function can also be expressed as

$$n(p - \frac{1}{m}\sum_{j=1}^{m}\sum_{s=1}^{p}\eta_{js}^2) = n(p - \sum_{s=1}^{p}\gamma_s), \tag{2.28}$$

where the quantities $\gamma_s = \frac{1}{m}\sum_{j=1}^{m}\eta_{js}^2$, $s = 1, ..., p$ called the eigenvalues, correspond to the average of the discrimination measures, and give a measure of the fit of the Homals solution in the sth dimension.

2.6.4 Properties of HOMALS

Some of the basic properties of HOMALS are listed below.

1. Category quantifications and object scores are represented as points in joint space.
2. Category points are the center of gravity of the object points that share the same category.
3. A variable discriminates better to the extent that the category points are farther apart.
4. If a category applies uniquely to only a single object, then the object point and that category point will coincide.
5. Category points with low marginal frequencies will be located further away from the origin of the joint space, whereas categories with high marginal frequencies will be located closer to the origin.
6. Objects with a 'unique' profile will be located further away from the origin of the joint space, whereas objects with a profile similar to the 'average' one will be located closer to the origin.
7. The category quantifications of each variable j have a weighted sum over categories equal to zero. This follows from the employed normalization of the object scores, since $u^{'}D_j\hat{Y}_j = u^{'}D_jD_j^{-1}G_j^{'}\hat{X} = u^{'}G_j^{'}\hat{X} = u^{'}\hat{X} = 0$.

8. The Homals solutions are nested. This means that if one requires a p_1-dimensional Homals solution and then a second p_2 greater than p_1 dimensional solution, then the first p_1 dimensions of the later solution are identical to the p_1-dimensional solution.

9. The solutions for subsequent dimensions are ordered. This means that the first dimension has the absolute maximum eigenvalue. The second dimension has the next maximum eigenvalue subject to the constraint that $X(.,2)$ is uncorrelated to $X(.,1)$, and so forth.

10. The solutions for the object scores are uncorrelated. However, the solutions for the quantifications need not necessarily be uncorrelated.

11. The solution is invariant under rotations of the object scores and of the category quantifications. To see this, we select a different basis for the column space of the object scores X; that is, let $X^\sharp = X \times R$, where R is a rotation matrix satisfying $R^{'}R = RR^{'} = I_p$. We then get that $Y_j^\sharp = D_j^{-1}G_j^{'}X^\sharp = \hat{Y}_jR$. Thus, the axes of the joint space can not be uniquely identified.

2.6.5 Homogeneity Analysis as an Eigenvalue and Singular Value Decomposition Problem

Homogeneity Analysis is appealing in the sense that its minimization problem can be treated as an eigenvalue problem. If we substitute the optimal $\hat{Y}_j = D_j^{-1}G_j^{'}X$ for given X, in the loss function, the loss function in (2.20) can be given as

$$\begin{aligned}
\sigma(X;*) &= \frac{1}{m}\sum_{j=1}^{m} tr[(X - G_jD_j^{-1}G_j^{'}X)^{'}(X - G_jD_j^{-1}G_j^{'}X)] \\
&= \frac{1}{m}\sum_{j=1}^{m} tr(X^{'}X - X^{'}G_jD_j^{-1}G_j^{'}X), \qquad (2.29)
\end{aligned}$$

where the symbol $*$ has replaced the argument over which the loss function is minimized. Let $P_j = G_jD_j^{-1}G_j^{'}$ denote the orthogonal projector on the subspace spanned by the columns of the indicator matrix G_j. Equation (2.29) can be rewritten as

$$\begin{aligned}\sigma(X;*) &= \frac{1}{m}\sum_{j=1}^{m} tr[(X-P_jX)^{'}(X-P_jX)] \\ &= \frac{1}{m}\sum_{j=1}^{m} tr(X^{'}X - X^{'}P_jX). \end{aligned} \tag{2.30}$$

Let P_* be the average of the m projectors. The equation (2.30) together with the normalization constraints in (2.21) and (2.22) gives that maximizing (2.30) comes to maximizing $tr(X^{'}\zeta P_*\zeta X)$, where $\zeta = I - uu^{'}/u^{'}u$ is a centering operator that leaves ζX in deviations from its column means. The optimal X corresponds to the first p eigenvectors of the matrix $\zeta P_*\zeta$. We can write the minimum loss as

$$\sigma(*;*) = m(p - \sum_{s=1}^{p}\lambda_s), \tag{2.31}$$

where $\lambda_s,\ s = 1, ..., p$ are the first p eigenvectors of P_*. Therefore, the minimum loss of homogeneity analysis is a function of the p largest eigenvalues of the average projector P_*.

The solution for optimal X can be obtained by the singular value decomposition of

$$m^{-1/2}\zeta G D^{-1/2} = U \wedge V \tag{2.32}$$

where the left-hand side is the super-indicator matrix in deviations from column means and corrected for marginal frequencies. The optimal X corresponds to the first p columns of the matrix U (the first p left singular vectors).

2.6.6 Homogeneity Analysis with Missing Data

In the presence of missing data, the loss function then becomes

$$\sigma(X; Y_1, ..., Y_j) = m^{-1}\sum_{j=1}^{m} tr[(X - G_jY_j)^{'}M_j(X - G_jY_j)] \tag{2.33}$$

subject to the normalization constraint $X^{'} M_* X = mnI_p$ and $u^{'} M_* X = 0$ where M_j and M_* are explained in detail in the categorical data coding section.

2.6.7 Relationship between HOMALS and Linear PCA

HOMALS is related to linear PCA in the following way. We start with having a look at the first HOMALS dimension. Let Q_1 be the optimally scaled data matrix and let the correlation matrix between the transformed variables in Q_1 be denoted by R_1. We assume that the columns of Q_1 are unit normalized. Hence, we write $R_1 = Q_1^{'} Q_1$. We write the singular value decomposition of Q_1 as $Q_1 = K_1 \wedge_1 L_1^{'}$ and the eigenvalue decomposition of R_1 as $R_1 = K_1 \wedge_1^2 L_1^{'}$. It can be shown that the normalized object scores in the first HOMALS dimension x_1 are proportional (with respect to a factor $n^{1/2}$) to the normalized component scores on the basis of Q_1, which are obtained by taking k_1. The discrimination measures are equal to the squares of the component loadings in the first PCA dimension, which are obtained by $a_1 = \lambda_1 l_1$. Refer [Gifi, 1989] Chapter 3 for the proof.

2.6.8 Relationship between HOMALS and Chi-Square

Let T be an $i \times j$ contingency table, whose entries t_{ij} give the frequencies with which row category i occurs together with column category j. Let $r = Tu$ denote the vector of row marginals, $c = T^{'} u$ the vector of column marginals and $n = u^{'} c = u^{'} r$ the total number of observations. Let $D_r = \text{diag}(r)$ be the diagonal matrix containing the elements of vector r and $D_c = \text{diag(c)}$ the diagonal matrix containing the elements of vector c. The χ^2-distances between rows i_1 and i_2 of table T is given by

$$\delta^2(i_1, i_2) = n \sum_{j=1}^{m} \frac{(t_{i_1 j}/r_{i_1} - f_{i_2 j}/r_{i_2})^2}{c_j} \tag{2.34}$$

Equation (2.34) shows that $\delta^2(i_1, i_2)$ is a measure for the difference between the profiles of rows i_1 and i_2.

To derive the coordinates X of the row categories of table T in the new Euclidean space, we consider the singular value decomposition of the matrix of the observed frequencies minus the expected frequencies corrected for row and column marginals

$$D_r^{-1/2}(F-E)D_c^{-1/2} = U \wedge V^{'} \tag{2.35}$$

where $E = rc^{'}/n$. The optimal scores X are then given (after normalization) by

$$X = n^{1/2}D_r^{-1}U \tag{2.36}$$

so that, $X^{'}D_rX = nI$ and $u^{'}D_rX = 0$.

Now consider the super-indicator matrix G. It resorts to the singular value decomposition of the matrix

$$\begin{aligned} m^{-1/2}(G - \frac{m}{mn}Guu^{'})D^{-1/2} &= m^{-1/2}\zeta GD^{-1/2} \\ &= U \wedge V \end{aligned} \tag{2.37}$$

which is identical to equation (2.32). This shows that homogeneity analysis could also be viewed as approximating the χ^2-distances between rows of the super-indicator matrix. This is due to the fact that the row marginals of the super-indicator matrix are all equal to m.

Chapter 3

Multivariate Binary Mixture-Model for Cluster Analysis

3.1 Introduction

A binary variable is one that can take values 0 and 1 which indicates the absence and presence of that variable respectively. Let P be a population (multivariate binary data) consisting of n objects where each object is an observation on each of the J binary variables. Cluster Analysis, [Aderberg, 1973], is a technique of grouping these n objects from population P into one or more groups such that the objects within each group are similar and the objects between each group are quite dissimilar. Multivariate binary data arises in most of the disciplines such as chemistry, pharmacology, ecology, genetics, and social science ([Larsen and Liu, 2005]).

A finite mixture model is one that is comprised of two or more finite probability density functions ([Titterington et al., 1985], [McLachlan and Peel, 2000], [Lindsay, 1995]). Finite mixtures of multivariate Bernoulli distributions have been extensively used in diverse fields. In finite mixture modeling, most of the emphasis in the literature had been on Gaussian mixture models ([Dasgupta, 1999], [Dasgupta and Schulman, 2000], [Arora and

Kannan, 2001]) and little attention has been on Bernoulli mixture models since the class of finite mixtures of multivariate Bernoulli distributions is known to be non-identifiable i.e., different values of the mixture parameters can correspond to exactly the same probability distribution. [Carreira-Perpinan and Renals, 2000] gave an empirical support to the fact that estimation of this class of mixtures can still produce meaningful results in practice, thus lessening the importance of the identifiability problem.

[Carreira-Perpinan, 2001] discussed the clustering of finite mixtures of Bernoulli distributions where the selection of optimal number of mixtures is based on the minimum lack of fit criteria. We show that the minimum lack of fit criteria is not just sufficient in determining the optimal number of mixtures. We show that the information criteria, ICOMP ([Bozdogan, 1987]) together with the lack of fit can determine the optimal number of mixtures in this case.

High dimensional data has been a problem by many researches in cluster analysis. It might be computationally expensive and convergence to the optimal parameter values might be time consuming. Moreover, not all predictors in the data might be needed for classification into the target number of mixtures. Selecting the optimal number of predictors from such a large number of predictors might be a challenging problem. We address the problem of high dimensional binary data using a genetic algorithm ([Goldberg, 1989], [Holland, 1992], [Forrest, 1993], [Srinivas and Patnaik, 1994]).

This chapter is organized as follows. Section 2 gives a brief background on univariate and multivariate Bernoulli distribution and their respective moments. Finite mixture-model of multivariate Bernoulli distributions is discussed in section 3. A brief explanation of information complexity (ICOMP) for the mixture case and the reason for using genetic algorithm for high dimensional binary data is given in section 4 and section 5 respectively.

3.2 Background

We review some of the concepts of Bernoulli distribution ([Carreira-Perpinan, 2001]) in this section.

3.2.1 Univariate Bernoulli Distribution

Definition

The Bernoulli distribution is a discrete distribution having two possible outcomes $X = 0$ and $X = 1$ where $X = 1$ is called success and it occurs with probability p and $X = 0$ is called failure and it occurs with probability $q = 1 - p$ where $0 < p, q < 1$. The probability function of a univariate Bernoulli distribution is given by

$$P(x) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \end{cases} \tag{3.1}$$

The above function can also be written as

$$P(x) = p^x (1-p)^{1-x} \tag{3.2}$$

Moments

The moments of a univariate Bernoulli distribution of parameter p are given by

$$\begin{align} Mean : \mu &= p \tag{3.3} \\ Variance : \sigma^2 &= p(1-p) \tag{3.4} \end{align}$$

3.2.2 Multivariate Bernoulli Distribution

Definition

A D-variate Bernoulli distribution of parameter $\mathbf{p} = (p_1, \ldots, p_D)^T \in [0,1]^D$, $B_D(p)$, is defined as

$$P(t;p) = \prod_{d=1}^{D} p_d^{t_d}(1-p_d)^{1-t_d} = \prod_{d=1}^{D} P(t_d|B(p_d) \tag{3.5}$$

where $B(p_d)$ is a Bernoulli distribution of parameter $p_d, d = 1, \ldots, D$. Thus, the D-variate Bernoulli distribution is equivalent to D independent Bernoulli distributions.

Moments

The moments of the D-variate Bernoulli distribution of parameter p are given as:

$$\begin{aligned} mean : \mu &= \mathbf{p} & (3.6) \\ covariance : \Sigma &= diag(p_d(1-p_d)) & (3.7) \end{aligned}$$

3.3 Finite Mixture-Model of Multivariate Bernoulli Distributions

A mixture of M D-variate Bernoulli distribution $B_D(p_d), \ldots, B_D(p_M)$ is defined as:

$$p(t; \{\pi_m, p_m\}_{m=1}^{M}) = \sum_{m=1}^{M} \pi_m p(t|m) \tag{3.8}$$

where the mixing proportions π_m satisfy $0 < \pi_m < 1$ for $m = 1, \ldots, M$ and $\sum_{m=1}^{M} \pi_m = 1$ and the component distributions are D-variate Bernoulli distributions, $t|m \sim B_D(p_m)$. In case of $M = 1$, choose $\pi_1 = 1$.

The moments for the mixture of M D-variate Bernoulli distributions are given as:

$$\begin{aligned} mean : \mu &= \sum_{m=1}^{M} \pi_m \mu_m \qquad (3.9) \\ covariance : \Sigma &= \sum_{m=1}^{M} \pi_m E_{p(t|m)}\{tt^T\} - \mu\mu^T \\ &= \sum_{m=1}^{M} \pi_m (\Sigma_m + \mu_m \mu_m^T) - \mu\mu^T, \qquad (3.10) \end{aligned}$$

where for $m = 1, \ldots, M, \mu_m = p_m$ and $\Sigma_m = diag(p_{md}(1 - p_{md}))$ are the component means and covariance matrices, respectively. Expanding Σ obtains:

$$\begin{aligned} (\Sigma)_{de} &= \sum_{n>m} \pi_m \pi_n (p_{md} - p_{nd})(p_{me} - p_{ne}) \qquad (3.11) \\ (\Sigma)_{dd} &= \mu_d(1 - \mu_d). \qquad (3.12) \end{aligned}$$

Since Σ is no longer diagonal, a mixture of multivariate Bernoulli distribution can account for correlations between variables.

3.3.1 Log likelihood of multivariate Bernoulli distribution

Let M be a fixed number of components. Let $\pi = (\pi_1, \ldots, \pi_M)^T$ and $P = (p_1, \ldots, p_M)$. The log likelihood of the parameters $\{\pi, P\}$ given a sample $\{t_n\}_{n=1}^N$ is

$$\begin{aligned} L(\pi, P) &= \sum_{n=1}^{N} ln(p(t_n; \pi, P)) \\ &= \sum_{n=1}^{N} ln(\sum_{m=1}^{M} \pi_m \prod_{d=1}^{D} p_{md}^{t_{nd}} (1 - p_{md})^{1-t_{nd}}). \qquad (3.13) \end{aligned}$$

3.3.2 Maximum likelihood parameter estimation

The maximum likelihood parameters are estimated by using an EM algorithm. The gradient density of the log likelihood of multivariate Bernoulli distribution is given by

$$\begin{aligned}
\frac{\partial L}{\partial \pi_m} &= \frac{1}{\pi_m}\sum_{n=1}^{N} p(m|t_n;\pi,P) - N \\
&\quad m = 1,\ldots,M && (3.14)\\
\frac{\partial L}{\partial p_{md}} &= \frac{1}{p_{md}(1-p_{md})}\sum_{n=1}^{N} p(m|t_n;\pi,P)(t_{nd}-p_{md}) \\
&\quad m = 1,\ldots,M \quad d = 1,\ldots,D && (3.15)
\end{aligned}$$

where

$$\begin{aligned}
p(m|t_n;\pi,P) &= \frac{p(t_n|m;\pi,P)p(m)}{\sum_{m'=1}^{M} p(t_n|m';\pi,P)p(m')} \\
&= \frac{\pi_m \prod_{d=1}^{M} p_{md}^{t_{nd}}(1-p_{md})^{1-t_{nd}}}{\sum_{m'=1}^{M} \pi_{m'} \prod_{d=1}^{D} p_{m'd}^{t_{nd}}(1-p_{m'd})(1-t_{nd})}
\end{aligned}$$

The basic equations for the derivations of the EM algorithm for finite mixture of multivariate Bernoulli distributions are given below:

E step: computation of the responsibilities using the above $p(m|t_n;\pi,P)$ from the current parameter estimates $\{\pi^{(\tau)},P^{(\tau)}\}$ at iteration $\tau, p(m|t_n;\pi^{(\tau)},P^{(\tau)})$.

M step: re-estimation of $\{\pi^{(\tau+1)},P^{(\tau+1)}\}$;

$$\begin{aligned}
\pi_m^{(\tau+1)} &= \frac{1}{N}\sum_{n=1}^{N} p(m|t_n;\pi^{\tau},P^{\tau}) \\
p_m^{(\tau+1)} &= \frac{1}{N\pi_m^{(\tau+1)}}\sum_{n=1}^{N} p(m|t_n;\pi^{(\tau)},P^{(\tau)})t_n
\end{aligned}$$

The sequence of parameters obtained for $\tau = 0,1,2,\ldots$ by iterating between the E and M steps from any starting point $\{\pi^{(0)},P^{(0)}\}$ produces a monotonically increasing sequence of values for the log-likelihood [Dempster et al., 1977].

A common problem of estimation in mixture distributions is that of singularities, that is, points in parameter space whose log-likelihood tends to positive infinity. Such singularities are undesirable because they give rise to degenerate distributions. Fortunately, the log-likelihood surface of a finite mixture of multivariate Bernoulli distributions has no singularities of value $+\infty$. The reason is that both the log-likelihood and its gradient are bounded above in the whole parameter space, including its boundaries. This means that estimation by the above EM algorithm from any non pathological starting point, which is always possible by choosing p_{md} in (0, 1), will always lead to a proper stationary point of the log-likelihood.

3.4 Information Complexity in Binary Mixture-Modeling

The very first step in mixture modeling is to determine the number of mixtures that actually fit the data. Let $K = 1, \ldots, K_M$ be the number of mixtures that can be used to fit the distribution. [Bozdogan, 1987] gave several guidelines for determining the limit of K_M.

$$K_M < ceil(\frac{2N}{(D+1)(D+2)}) \tag{3.16}$$

$$K_M \cong ceil(\sqrt{\frac{N}{2}}) \tag{3.17}$$

$$K_M = ceil(log_2 N) \tag{3.18}$$

Once K_M is determined, we need to find the best number of mixtures K from the K_M different arrangements of the data. Information complexity helps in determining the best number of mixtures from various arrangements of the data. The criterion for choosing the best model for the data is to choose the model that has the lowest information criteria value.

The usual penalty term used in AIC is $2k$ where k is the total number of parameters in the model. Therefore,

$$AIC = -2Loglikelihood + 2k$$

But in the case of mixture models, the penalty term is $3k$. In the D-Bernoulli distribution case with K mixtures, we have $K \times D$ probability values to estimate and $(K-1)$ mixture parameters to estimate. Therefore, the estimated k in this case would be:

$$k = K \times D + (K-1)$$

where K is the number of mixtures of Bernoulli distribution and D is in D-variate Bernoulli distribution for each mixture.

Therefore, for the mixture of Bernoulli distributions, we can write AIC and SBC as

$$AIC = -2\sum_{n=1}^{N} ln(\sum_{m=1}^{M} \pi_M \prod_{d=1}^{D} p_{md}^{t_{nd}}(1-p_{md})^{1-t_{nd}}) + 3k \tag{3.19}$$

$$SBC = -2\sum_{n=1}^{N} ln(\sum_{m=1}^{M} \pi_m \prod_{d=1}^{D} p_{md}^{t_{nd}}(1-p_{md})^{(1-t_{nd})}) + log(N)k \tag{3.20}$$

ICOMP for the normal models is given as

$$ICOMP(\hat{Cov}) = -2Loglikelihood + 2C_{1F}(\hat{Cov}) \tag{3.21}$$

where $\hat{Cov}$ is the estimated covariance matrix of the mixture density and $C_{1F}(\hat{Cov})$ is given as:

$$C_{1F}(\hat{Cov}) = \frac{s}{2}log(\frac{trace(\hat{Cov})}{s}) - \frac{1}{2}log(|\hat{Cov}|), \qquad s = rank(\hat{Cov}) \tag{3.22}$$

The $\hat{Cov}$ in this case, is given by

$$\begin{aligned} \hat{Cov} &= \hat{F}_\pi^{-1} \\ &= \begin{bmatrix} \hat{F}_\pi^{-1} & 0 & \dots & 0 \\ 0 & \hat{F}_1^{-1} & \dots & 0 \\ & & \ddots & \\ 0 & \dots & \dots & \hat{F}_M^{-1} \end{bmatrix} \end{aligned} \tag{3.23}$$

For $m = 2$,

$$\begin{aligned}\hat{F}^{-1}(\hat{\pi}) &= \begin{bmatrix} \frac{1}{\hat{\pi}_1} & 0 \\ 0 & \frac{1}{\hat{\pi}_2} \end{bmatrix} \\ &= \hat{Cov}(\hat{\pi}) \qquad (3.24)\end{aligned}$$

$$\hat{F}^{-1}(1) = \begin{bmatrix} \hat{\Sigma}_1 & 0 \\ 0' & (\frac{2}{n_1})D_p^+(\hat{\Sigma}_1 \otimes \hat{\Sigma}_1)D_p^{+'} \end{bmatrix} \qquad (3.25)$$

$$\hat{F}^{-1}(2) = \begin{bmatrix} \hat{\Sigma}_2 & 0 \\ 0' & (\frac{2}{n_2})D_p^+(\hat{\Sigma}_2 \otimes \hat{\Sigma}_2)D_p^{+'} \end{bmatrix} \qquad (3.26)$$

In the case of mixture of D-Bernoulli distributions, the penalty term in the ICOMP is even higher than the normal case. It is given by

$$ICOMP(\hat{Cov}) = -2Loglikelihood + 2C_{1F}(\hat{Cov}) + 3k \qquad (3.27)$$

In the mixture case, $ICOMP_{PEU\_LN}$ and $ICOMP_{PEU\_Misspec}$ is given by,

$$\begin{aligned}ICOMP_{PEU\_LN}(\hat{Cov}) &= -2\sum_{n=1}^{N} ln(\sum_{m=1}^{M} \pi_m \prod_{d=1}^{D} p_{md}^{t_{nd}}(1-t_{nd})) \\ &\quad +3k + log(N)C_{1F}(\hat{Cov}) \qquad (3.28)\end{aligned}$$

$$\begin{aligned}ICOMP_{PEU\_Misspec}(\hat{Cov}) &= -2\sum_{n=1}^{N} ln(\sum_{m=1}^{M} \pi_m \prod_{d=1}^{D} p_{md}^{t_{nd}}(1-t_{nd})) \\ &\quad +2\frac{Nk}{N-k-2} + 2C_{1F}(\hat{Cov}) \qquad (3.29)\end{aligned}$$

3.5 High Dimensional Binary Data

If data contains many predictors (generally greater than 10), the EM algorithm in the above Bernoulli case might be computationally expensive and convergence to the optimal parameter values might be time consuming. Moreover, not all predictors in the data might be needed for classification into the target number of mixtures. Selecting the optimal number of predictors from such a large number of predictors might be a challenging problem.

One quick solution to this problem might be using the all possible subsets approach. This approach has a drawback. For high dimensional data, the number of all possible subsets is large and hence needs a lot of computation expense. For example, if the number of predictors in the data set are 10. The number of all possible subsets is $2^{10} = 1024$. Other conventional optimization techniques such as steepest ascent, conjugate gradient etc., might be restricted to local optima. Hence, we need to have an efficient optimization procedure which does not restrict itself to local optima. We use the concept of genetic algorithm to search for the optimal or near optimal solution from this vast solution space.

Chapter 4

Mixed Data

4.1 Introduction

4.1.1 Definition

Mixed Data can be defined as a combination of quantitative and qualitative data variables. Real world data are not all quantitative or not all qualitative. Mostly, they are a combination of quantitative and qualitative data.

4.1.2 Example

The data shown in table 4.1 are a perfect example of mixed data. The variables age, height and weight are continuous. The variables gender and smoker are nominal and the variable CARDIAC (test for a cardiac arrest) is ordinal.

Table 4.1: Example of a Mixed Data

Age	Gender	Height(in cm)	Weight(in lbs)	Smoker	CARDIAC
32	Male	176.2	192.4	0	Negative
35	Female	164.3	154.2	0	Negative
48	Male	175.3	162.7	1	Positive
36	Male	180.1	200.6	1	Positive
52	Female	154.9	143.4	0	Negative
63	Female	157.2	142.9	0	Negative
53	Male	165.5	176.2	1	Positive
47	Male	173.6	192.4	0	Negative
35	Female	164.1	153.4	0	Negative
41	Male	172.9	220.7	1	Positive

4.2 Problems with Mixed Data

Researchers in statistical data analysis usually face problems if the data are of the mixed type. Most of the univariate and multivariate statistical concepts deals with continuous or categorical data but not mixed data. The traditional statistical techniques performed with the presence of a qualitative variable in the data set containing other quantitative variables might not give accurate results. If qualitative variables are only a few when compared to quantitative variables, the usual practice followed by researchers is to just leave the qualitative variable intact (treat them as continuous variables) and analyze the data. Another option is to drop the qualitative variables and analyze only the quantitative variables in the data. For instance, in multiple regression, where the response is a continuous variable, by removing the qualitative variables from the model we lose significant amount of knowledge regarding the effects that a qualitative variable has on the continuous response variable. The other option is to use dummy variables. Similar is the case when a data set has most of the variables qualitative and a few variables quantitative.

We illustrate a procedure described in [Lee, 2007]. Consider an ordered categorical variable with a five-point scale 1, 2, 3, 4, 5 corresponding to the answer on the opinion of a policy. The description for each of the scales are 'strongly disagree', 'disagree', 'no opinion', 'agree' and 'strongly agree'. One common approach is to treat the assigned integers as continuous data from a normal distribution. This approach may not lead to serious problems if the histogram of the observations is symmetrical and with the highest frequency at the center. This is the situation where most subjects choose the category 'no opinion'. To claim multivariate normality of the observed variables, we need to have most subjects choosing the middle category, for example 'no opinion' or 'no change', in all the corresponding items. However, for an interesting item in the questionnaire, most subjects would be likely to select categories at both ends, for example, 'strongly agree (strongly disagree)' or 'agree (disagree)'. Hence, in practice, histograms corresponding to most variables are either skewed or bi-modal. Clearly, routinely treating ordered categorical variables as normal may lead to erroneous conclusions ([Lee et al., 90a], [Lee et al., 90b], [Lee et al., 1995]).

4.3 How to handle Mixed Data

[Lee, 2007] describes a better approach for assessing discrete data. The approach is to treat them as observations that are coming from a hidden continuous normal distribution with a threshold specification. Suppose for a given data set, the proportions of 1, 2, 3, 4 are 0.05, 0.05. 0.4 and 0.5, respectively. If we make a histogram for this discrete data it would be highly skewed to the right. The threshold approach for analyzing this highly skewed discretized variable is to treat the ordered categorical data as manifestations of an underlying normal variable y. The exact continuous measurements of y are not available, but are related to the observed ordered categorial variable z such as follows: for k = 1, 2, 3, 4

$$\text{z = k if } \alpha_{k-1} < y \leq \alpha_k;$$

where $-\infty = \alpha_0 < \alpha_1 < \alpha_2 < \alpha_3 < \alpha_4 = \infty$, and α_1, α_2 and α_3 are thresholds. Then the ordered categorical observations can be captured by N(0,1) with appropriate thresholds. As α_2 - α_1 can be different from α_3 - α_2, unequal-interval scales are allowed. Hence, this threshold approach allows flexible modeling. As it is related to a common normal distribution, it also provides easy interpretation of the parameters. It should be noted that the ad hoc integral values, here k = 1, 2, 3, 4, are solely used to represent the category; only their frequencies are important in the statistical analysis.

Mixed Data can be handled by transforming the qualitative data into quantitative form. If all the data variables are quantitative, the usual classical multivariate analysis (MVA) can be performed. In this work, we use the Gifi transformation, [Gifi, 1989], to transform the qualitative variables to quantitative form. Detailed description about the Gifi transformation is given in the Literature Review Chapter.

Chapter 5

Gifi System – Applications

In this chapter, we analyze the data in the Gifi space with several traditional multivariate statistical methods such as multiple regression, binary logistic regression, multivariate regression, multivariate logistic regression, discriminant analysis and cluster analysis. For each application, we introduce two algorithms that can be used to analyze the data in the Gifi space. One algorithm (OSM - Optimal Scaling Method) optimally scales the categorical variables in the Gifi space, thus making the data set purely continuous in the Gifi space. Hence in the optimally scaled version, the p-dimensional categorical variables are transformed to a p-dimensional continuous variables. The other algorithm (LCM - Linear Combination Method) does a linear combination of the categories of the categorical variables thus making it a 1-dimensional continuous space. Hence, in the second version, a p-dimensional categorical variables are transformed to a 1-dimensional continuous space. The OCM might be useful when there are a few categorical variables in the data set whereas the LCM might be useful when the dimension of the categorical variables is very large.

5.1 Data Transformation

Consider a mixed data set D_0 consisting of variables $x_1, x_2, x_3, x_4, x_5, y$ and $n = 100$ observations. Suppose that the variables x_1, x_2, x_4 are categorical and the rest are continuous. The input to the Gifi system will be the data set D_1, where D_1 contains data on the variables x_1, x_2, x_4. After the transformation using LCM, the original data set D_0 can be represented by a transformed data set D_3, where D_3 contains a linear combination of the weights of the categories of x_1, x_2, x_4 in 1-dimension in addition to the original linear values of x_3, x_5, y. Therefore, after the transformation the three dimensional categorical space becomes a one dimensional continuous space. Hence the transformed Gifi space would be of size 100×4. We use the normal scores algorithm adapted to heterogeneous variances for transforming the categorical space to a continuous space. Hence, in this example, the transformed Gifi space would be of size 100×4. After the transformation using OCM, the original data set D_0 can be represented by a transformed data set D_3 where $D_3 = \{G_1 \times y_1, G_2 \times y_2, x_3, G_4 \times y_4, x_5, y\}$ where G_i is the indicator matrix for a categorical variable i and y_i is the set of corresponding weights for the categories of a categorical variable i. Here also we use the normal scores algorithm adapted to heterogeneous variances for transforming the categorical space to a continuous space. Hence, in this example, the transformed Gifi space would be of size 100×6.

5.2 Regression

Regression analysis is a statistical methodology to estimate the relationship of a response (or a dependent) variable to a set of predictor (or independent) variables. This technique can be performed on one or more than one dependent variable(s). Regression analysis on one dependent variable and one or more independent variables is known as multiple regression. A simple regression, having one dependent variable and one independent variable, is a special case of multiple regression. Regression analysis on a data set having more than one dependent variable and one or more independent variable(s) is known as multivariate regression.

Most of the work in the literature had been on linear regression analysis and less emphasis had been on nonlinear regression analysis. In the simple case, a linear regression analysis is about fitting a straight line,

$$y = \beta_0 + \beta_1 x, \tag{5.1}$$

to a set of paired data $\{(x_i, y_i), i = 1, 2, ..., n\}$ on two numerical variables x and y. The usual linear regression techniques generate good models when the data is purely continuous. If data contains mostly categorical variables, these techniques fail to generate good models. In this work, we use the Gifi transformation, [Gifi, 1989], on the non linear data and apply the usual linear regression analysis on the transformed linear data. In this chapter, we first explain the implementation of the Gifi system on a mixed data for multiple regression and then explain the implementation of the Gifi system on a mixed data for multivariate regression.

5.2.1 Multiple Regression

Methodology

After the transformation, in the Gifi space, the data set is no more of the mixed type. It would be purely in continuous form. Hence, we can apply the usual linear multiple regression technique on the transformed mixed data.

In multiple regression we fit a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon, \tag{5.2}$$

where $x_1, x_2, ..., x_k$ are k predictor variables, $\beta_0, \beta_1, ..., \beta_{k+1}$ are $k+1$ unknown parameters and ϵ the error term.

The multiple regression model 5.2 and the formulas for its estimation can be presented in a compact form if we use matrix notation [Tamhane and Dunlop, 2003]. Let

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}$$

where Y is a $n \times 1$ vector of the random variables $Y_i^{\prime}s$, y is a vector of observed values, ϵ is a vector of random errors and X is a $n \times (k+1)$ matrix of the values of predictor variables. The first column of X of all 1's corresponds to the constant term β_0 in the model 5.2. Also let,

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ . \\ . \\ . \\ \hat{\beta}_n \end{bmatrix}$$

where β and $\hat{\beta}$ are the $(k+1) \times 1$ vectors of unknown parameters and their least squares estimates, respectively.

Using this notation, the model in 5.2 can be written as

$$Y = X\beta + \epsilon \tag{5.3}$$

The matrix notation to obtain the least squares estimates is represented as

$$X^{'}X\beta = X^{'}y \tag{5.4}$$

If the inverse of the matrix $X^{'}X$ exists, then the solution is given by

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}y \tag{5.5}$$

We assume that the errors are normally distributed. Let $\hat{\sigma^2}$ be the variance of the residuals. The negative loglikelihood in the normal case is given by

$$-logL(\hat{\beta}, \hat{\sigma^2}) = \frac{n}{2}log(2\pi) + \frac{n}{2}log(\hat{\sigma^2}) + \frac{n}{2} \tag{5.6}$$

where the parameter $\hat{\sigma^2}$ is estimated by

$$\hat{\sigma^2} = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta}), \tag{5.7}$$

Information Criteria

The various information criteria's for the multiple regression case are briefly described in this section. Let k be the number of parameters in the model. For now, we assume a normal distribution on the residuals.
AIC, [Akaike, 1973], is given by

$$AIC = nlog(2\pi) + nlog(\hat{\sigma^2}) + n + 2k \tag{5.8}$$

CAIC, [Bozdogan, 1987], is given by

$$CAIC = nlog(2\pi) + nlog(\hat{\sigma^2}) + n + k(log(n) + 1) \tag{5.9}$$

The IFIM (inverse fisher information matrix) version of ICOMP ([Bozdogan, 1987], [Bozdogan, 1988], [Bozdogan, 90a], [Bozdogan, 90b], [Bozdogan, 2004]) in its Frobenius norm characterization is given by

$$ICOMP_{1F}(IFIM) = nlog(2\pi) + nlog(\hat{\sigma^2}) + n + 2C_{1F}(\hat{F}^{-1}) \tag{5.10}$$

where

$$\hat{F}^{-1} = \begin{bmatrix} \frac{\hat{\sigma^2}}{\sum_{i=1}^{n} x_i^2} & 0 \\ & \\ 0 & \frac{2\hat{\sigma^2}}{n} \end{bmatrix} \tag{5.11}$$

and

$$\begin{aligned} C_{1F}(\hat{F^{-1}}) &= \frac{s}{4}\frac{(1/s)tr(\hat{F^{-1}}\hat{F^{-1}}^{'}) - (\frac{tr(\hat{F^{-1}})}{s})^2}{(\frac{tr(\hat{F^{-1}})}{s})^2} \\ &= \frac{1}{4\bar{\lambda}^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda})^2 \cong C_1(\hat{F^{-1}}) \end{aligned} \tag{5.12}$$

To compute the ICOMP for the misspecification case, first we need to compute the outer product form of FIM. The estimated outer product form of FIM is given by

$$\hat{R}(\hat{\theta}) = \begin{bmatrix} \frac{n}{\hat{\sigma^2}} & \frac{nS_k}{2\hat{\sigma^3}} \\ & \\ \frac{nS_k}{2\hat{\sigma^3}} & \frac{n(K_t-1)}{2\hat{\sigma^4}} \end{bmatrix} \tag{5.13}$$

where S_k is the coefficient of skewness, $S_k = \frac{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^3}{\hat{\sigma^3}}$ and K_t is the coefficient of kurtosis, $K_t = \frac{\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^4}{\hat{\sigma^4}}$. $ICOMP_{MISSPEC}$ is given by

$$ICOMP_{MISSPEC} = nlog(2\pi) + nlog(\hat{\sigma^2}) + n + 2C_1(\hat{F}^{-1}\hat{R}(\hat{\theta})\hat{F}^{-1}) \tag{5.14}$$

SBC, [Schwarz, 1978], is given by

$$SBC = nlog(2\pi) + nlog(\hat{\sigma^2}) + n + log(n)k \tag{5.15}$$

Algorithm: Optimal Scaling Method

This algorithm fits a multiple regression model for a continuous response and a mixed set of predictors, X, and also selects the optimal predictors that explain most of the variation in the response.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Continuous Response Data : y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Run the Gifi transformation on the data X and optimally scale the categorical variables in the Gifi space. A categorical variable j is optimally scaled by multiplying its indicator matrix, G_j, with its optimal weight vector, y_j. Suppose if the data contains variables $[x_1, x_2, x_3, x_4, x_5]$. Let x_1 and x_4 be continuous and x_2, x_3, x_5 be categorical. Let G_2 and y_2 be the indicator matrix and the optimal weight vector for the categorical variable x_2 respectively. Similarly, G_3, G_5 and y_3, y_5 are the indicator matrices and optimal weight vectors of the categorical variables x_3 and x_5 respectively. Therefore, the data matrix in the Gifi space will be of the form $\{x_1, G_2 \times y_2, G_3 \times y_3, x_4, G_5 \times y_5\}$.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model in the Gifi space. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}.
    - Perform multiple regression with y as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else
    - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
    - Go to step 3

Algorithm: Linear Combination Method

This algorithm fits a multiple regression for a continuous response and a mixed set of predictors, X, and also selects the optimal predictors that explain most of the variation in the response.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Continuous Response Data : y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Transform the mixed data set X to a pure continuous space using Gifi transformation.
2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.
3. For each chromosome in the population
    - Build a new predictor data matrix, X_{new}. Since X_{new} might be a mixed data set, we split the X_{new} matrix into X_{con} and X_{cat} where X_{con} is the

data on the continuous predictors and X_{cat} is a 1-dimensional continuous data of the categorical predictors in the Gifi space. Hence X_{new} can be represented as $X_{new} = \left[\begin{array}{cc} X_{con} & X_{cat} \end{array} \right]$. Suppose, if the current chromosome selects x_1, x_2, x_3, x_4, x_5 where x_1, x_2, x_3, x_4, x_5 are a subset of the original set of predictors $x_1, ..., x_p$ where $p \geq 5$. Suppose x_1, x_3 are continuous and x_2, x_4, x_5 are categorical. We perform the Gifi transformation on x_2, x_4, x_5 and transform it to a 1-dimensional continuous space, X_{cat}. Since x_1, x_3 are continuous, $X_{con} = \left[\begin{array}{cc} x_1 & x_3 \end{array} \right]$. Therefore, $X_{new} = \left[\begin{array}{cc} X_{con} & X_{cat} \end{array} \right]$.

- Perform multiple regression with y as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

   - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
   - Go to step 2

5.2.2 Binary Logistic Regression

Methodology

Binary Logistic Regression (BLR) is a parametric method for regression when $Y_i \in 0, 1$ is binary [Wasserman, 2004]. For k-dimensional covariate X, the model is

$$\begin{aligned} p_i &\equiv p_i(\beta) \\ &\equiv P(Y_i = 1 | X = x) \\ &= \frac{e^{\sum_{j=1}^{k} \beta_j x_{ij}}}{1 + e^{\sum_{j=1}^{k} \beta_j x_{ij}}} \end{aligned} \tag{5.16}$$

or, equivalently,

$$logit(p_i) = \sum_{j=1}^{k} \beta_j x_{ij} \tag{5.17}$$

where

$$logit(p) = log(\frac{p}{1-p}) \tag{5.18}$$

The name logistic regression comes from the fact that $\frac{e^x}{1+e^x}$ is called the logistic function.

Because the Y_i's are binary, the data are Bernoulli:

$$Y_i | X_i = x_i \sim Bernoulli(p_i) \tag{5.19}$$

Hence the (conditional) likelihood function is

$$\zeta(\beta) = \prod_{i=1}^{n} p_i(\beta)^{Y_i}(1 - p_i(\beta))^{1-Y_i} \tag{5.20}$$

The MLE of $\hat{\beta}$ has to be obtained by maximizing $\zeta(\beta)$ numerically by the reweighted least squares algorithm.

Reweighted Least Squares Algorithm

Choose starting values $\hat{\beta}^0 = (\hat{\beta}_1^0, ..., \hat{\beta}_k^0)$ and compute p_i^0 using equation 5.16, for $i = 1, .., n$. Set $s = 0$ and iterate the following steps until convergence.

1. Set
$$Z_i = logit(p_i^s) + \frac{Y_i - p_i^s}{p_i^s(1-p_i^s)}, i = 1, ..., n$$

2. Let W be a diagonal matrix with (i,i) element equal to $p_i^s(1 - p_i^s)$.

3. Set
$$\hat{\beta}^s = (X^T W X)^{-1} X^T W Z$$

   This corresponds to doing a (weighted) linear regression of Z on X.

4. Set $s = s + 1$ and go back to the first step.

Information Criteria

The various information criteria's for the binary logistic regression case are briefly described in this section. Let k be the number of parameters in the model.

The loglikelihood for the binary logistic case is given by

$$log(\zeta(\beta)) = \sum_{i=1}^{n}[Y_i log(p_i(\beta)) + (1 - Y_i)log(1 - p_i(\beta))] \tag{5.21}$$

AIC, [Akaike, 1973], is given by

$$AIC(k) = -2log(\zeta(\beta)) + 2k \tag{5.22}$$

CAIC, [Bozdogan, 1987], is given by

$$CAIC(k) = -2log(\zeta(\beta)) + k(log(n) + 1) \tag{5.23}$$

SBC, [Schwarz, 1978], is given by

$$SBC(k) = -2log(\zeta(\beta)) + klog(n) \tag{5.24}$$

$ICOMP_{IFIM}$, ([Bozdogan, 1987], [Bozdogan, 1988], [Bozdogan, 90a], [Bozdogan, 90b], [Bozdogan, 2004]), is given by

$$ICOMP_{IFIM} = -2log(\zeta(\beta)) + 2C_1(\hat{F}^{-1}) \tag{5.25}$$

where $\hat{F}^{-1}$ is given by

$$\hat{F}^{-1} = \begin{bmatrix} \hat{\sigma}^2(X^TWX) & 0 \\ 0' & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} \tag{5.26}$$

and $\hat{\sigma}^2$ is estimated from

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(Y_i - p_i)^2}{p_i(1 - p_i)} \tag{5.27}$$

Algorithm: Optimal Scaling Method

This algorithm fits a binary logistic regression for a binary response, y and a mixed set of predictors, X, and also selects the optimal predictors that can best classify the data into two categories.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Binary Response Data : y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Run the Gifi transformation on the data set, X, and optimally scale the categorical variables in the Gifi space.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model in the Gifi space. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

   - Build a new predictor data matrix, X_{new}.
   - Perform binary logistic regression with the binary y as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

    - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
    - Go to step 3

Algorithm: Linear Combination Method

This algorithm fits a binary logistic regression for a binary response, y and a mixed set of predictors, X, and also selects the optimal predictors that can best classify the data into two categories.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Binary Response Data : y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Transform the mixed data set X to a pure continuous space using Gifi transformation.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}. Since X_{new} might be a mixed data set, we split the X_{new} matrix into X_{con} and X_{cat} where X_{con} is the data on

the continuous predictors and X_{cat} is a 1-dimensional continuous data of the categorical predictors in the Gifi space. Hence X_{new} can be represented as $X_{new} = \begin{bmatrix} X_{con} & X_{cat} \end{bmatrix}$.

- Perform binary logistic regression with the binary y as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N-1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

   - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
   - Go to step 2

5.2.3 Multi-class Logistic Regression

Methodology

The multi-class logistic regression is a parametric method for regression when the response, y, contains K classes where $K \geq 3$. We consider the covariate X to be p dimensional without the intercept term. We consider the K-th class to be the base class and fit $K-1$

logit transformations on the other $K-1$ classes.

$$\begin{aligned} log(\frac{P(G=1|X=x)}{P(G=K|X=x)}) &= \beta_{10}+\beta_1^T x \\ log(\frac{P(G=2|X=x)}{P(G=K|X=x)}) &= \beta_{20}+\beta_2^T x \\ &\vdots \\ log(\frac{P(G=K-1|X=x)}{P(G=K|X=x)}) &= \beta_{(K-1)0}+\beta_{(K-1)}^T x \end{aligned} \tag{5.28}$$

where $\beta_{10}, \beta_{20}, \ldots, \beta_{(K-1)0}$ are the coefficients of the intercept terms in the equation 5.28. Hence for any class pairs (k,l), we can write the logit transformation as

$$log(\frac{P(G=k|X=x)}{P(G=l|X=x)}) = \beta_{k0}-\beta_{l0}+(\beta_k-\beta_l)^T x \tag{5.29}$$

Therefore, the number of parameters in this model is given by

$$m = (K-1)\times(p+1) \tag{5.30}$$

Let us denote the parameter set by θ given by

$$\theta = \{\beta_{10}, \beta_1, \beta_{20}, \beta_2, \ldots, \beta_{(K-1)0}, \beta_{(K-1)}\} \tag{5.31}$$

The posterior probability that an observation x_i belongs to a class k is given by

$$P(G=k|X=x_i) = \frac{e^{\beta_{k0}+\beta_k^T x_i}}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_l^T x_i}} \qquad for \qquad k=1,...,K-1 \tag{5.32}$$

and for a class K it is given by

$$P(G=K|X=x_i) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_l^T x_i}} \tag{5.33}$$

The equations 5.32 and 5.33 ensures that the sum of the probabilities that an observation x_i belongs to classes $1, 2, ..., K$ equals 1.

Let there be N samples each having class $g_i, i = 1, 2, ..., N$. The conditional log-likelihood of the class labels is given by

$$l(\theta) = \sum_{i=1}^{N} log(P(G = g_i | X = x_i)) = \sum_{i=1}^{N} log(p_{g_i}(x_i; \theta)) \tag{5.34}$$

Since there are $K \geq 3$ classes, β is a $(K-1)(p+1)$ vector:

$$\beta = \begin{pmatrix} \beta_{10} \\ \beta_{11} \\ \vdots \\ \beta_{1p} \\ \beta_{20} \\ \vdots \\ \beta_{2p} \\ \vdots \\ \beta_{(K-1)0} \\ \vdots \\ \beta_{(K-1)p} \end{pmatrix} \tag{5.35}$$

Let

$$\bar{\beta_l} = \begin{pmatrix} \beta_{l0} \\ \beta_l \end{pmatrix} \tag{5.36}$$

Hence, the likelihood function is given by

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} log(p_{g_i}(x_i;\beta) \\ &= \sum_{i=1}^{N} log(\frac{e^{\bar{\beta}_{g_i}^T x_i}}{1+\sum_{l=1}^{K-1} e^{\bar{\beta}_l^T x_i}}) \\ &= \sum_{i=1}^{N} [\bar{\beta}_{g_i}^T x_i - log(1+\sum_{l=1}^{K-1} e^{\bar{\beta}_l^T x_i})] \end{aligned} \tag{5.37}$$

The first order derivative for the equation 5.37 is given by

$$\begin{aligned} \frac{\partial l(\beta)}{\partial \beta_{kj}} &= \sum_{i=1}^{N} [I(g_i = k) x_{ij} - \frac{e^{\bar{\beta}_k^T x_i}}{1+\sum_{l=1}^{K-1} e^{\bar{\beta}_l^T x_i}}] \\ &= \sum_{i=1}^{N} x_{ij}(I(g_i = k) - p_k(x_i;\beta)) \end{aligned} \tag{5.38}$$

where $I(.)$ is an indicator function which equals 1 when the argument is true and 0 otherwise.

The second order derivative for the equation 5.37 is given by

$$\frac{\partial^2 l(\beta)}{\partial \beta_{kj} \beta_{mn}} = -\sum_{i=1}^{N} x_{ij} x_{in} p_k(x_i;\beta)[I(k = m) - p_m(x_i;\beta)] \tag{5.39}$$

Reweighted Least Squares Procedure

In matrix form, we represent y as the concatenated indicator vector of dimension $N(K-1) \times 1$. It is given by

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{K-1} \end{pmatrix} \qquad y_k = \begin{pmatrix} I(g_1 = k) \\ I(g_2 = k) \\ \vdots \\ I(g_N = k) \end{pmatrix} \qquad 1 \leq k \leq K-1 \tag{5.40}$$

The fitted probabilities p is the concatenated vector of dimension $N(K-1)\times 1$. It is given by

$$p = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{K-1} \end{pmatrix} \qquad p_k = \begin{pmatrix} p_k(x_1;\beta) \\ p_k(x_2;\beta) \\ \vdots \\ p_k(x_N;\beta) \end{pmatrix} \qquad 1 \leq k \leq K-1 \tag{5.41}$$

$\tilde{X}$ is an $N(K-1)\times(p+1)(K-1)$ matrix. It is given by

$$\tilde{X} = \begin{pmatrix} X & 0 & \dots & 0 \\ 0 & X & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & X \end{pmatrix} \tag{5.42}$$

The weight matrix W is an $N(K-1)\times N(K-1)$ square matrix and is given by

$$W = \begin{pmatrix} W_{11} & W_{12} & \dots & W_{1(K-1)} \\ W_{21} & W_{22} & \dots & W_{2(K-1)} \\ \dots & \dots & \dots & \dots \\ W_{(K-1)1} & W_{(K-1)2} & \dots & W_{(K-1)(K-1)} \end{pmatrix} \tag{5.43}$$

where each submatrix $W_{km}, 1 \leq k, m \leq K-1$, is an $N \times N$ diagonal matix. When $k = m$, the i-th diagonal element in W_{kk} is given by

$$p_k(x_i;\beta^{old})(1 - p_k(x_i;\beta^{old})) \tag{5.44}$$

When $k \neq m$, the i-th diagonal element in W_{km} is given by

$$-p_k(x_i;\beta^{old})p_m(x_i;\beta^{old}) \tag{5.45}$$

The new β at each iteration is be given by

$$\beta^{new} = \beta^{old} + (\tilde{X}W\tilde{X})^{-1}\tilde{X}^T(y-p) \tag{5.46}$$

We can use $\beta = 0$ as one of the options as an initial starting point. One major problem with this approach is that convergence is not guaranteed in some cases.

Information Criteria

The various information criteria's for the binary logistic regression case are briefly described in this section. Let k be the number of parameters in the model.

AIC, [Akaike, 1973], is given by

$$AIC(k) = -2l(\beta)) + 2k \tag{5.47}$$

CAIC, [Bozdogan, 1987], is given by

$$CAIC(k) = -2l(\beta)) + k(log(n) + 1) \tag{5.48}$$

SBC, [Schwarz, 1978], is given by

$$SBC(k) = -2l(\beta)) + klog(n) \tag{5.49}$$

$ICOMP_{IFIM}$, ([Bozdogan, 1987], [Bozdogan, 1988], [Bozdogan, 90a], [Bozdogan, 90b], [Bozdogan, 2004]), is given by

$$ICOMP_{IFIM} = -2l(\beta)) + 2C_1(\hat{F}^{-1}) \tag{5.50}$$

where $\hat{F}^{-1}$ is given by

$$\hat{F}^{-1} = -H^{-1} \tag{5.51}$$

and H is the Hessian matrix of the equation 5.37. The Hessian matrix, H is a square matrix of the order $(K-1)(p+1) \times (K-1)(p+1)$. Each element of the Hessian matrix can be computed from the equation 5.39.

Algorithm: Multi-class Logistic Regression in Gifi space

The algorithm for the multi-class case is same as the one for the binary logistic regression case. But instead of using binary logistic regression we use multi-class logistic regression since there are $K \geq 3$ classes.

5.2.4 Multivariate Regression

Methodology

The methodology in the multivariate case is similar to the methodology in the multiple regression case since the transformed data set is purely continuous. Hence, we can apply the usual multivariate regression technique on the transformed mixed data.

Let Y be an $(n \times p)$ data matrix of n independent observations on p responses, X be the $(n \times q)$ design or model matrix of fixed known independent variables, B be the $(q \times p)$ matrix of coefficients to be estimated, and let E be the matrix of random errors. Then the multivariate linear regression model is given by

$$Y = XB + E, \tag{5.52}$$

where $q = k + 1$, $k =$ number of independent variables. In matrix notation,

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & ... & Y_{1p} \\ Y_{21} & Y_{22} & ... & Y_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ Y_{n1} & Y_{n2} & ... & Y_{np} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & ... & x_{1k} \\ 1 & x_{21} & x_{22} & ... & x_{2k} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}$$

$$B = \begin{bmatrix} \beta_{01} & \beta_{02} & ... & \beta_{0p} \\ \beta_{11} & \beta_{12} & ... & \beta_{1p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \beta_{k1} & \beta_{k2} & ... & \beta_{kp} \end{bmatrix} \qquad E = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & ... & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & ... & \varepsilon_{2p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \varepsilon_{n1} & \varepsilon_{n2} & ... & \varepsilon_{np} \end{bmatrix} \tag{5.53}$$

The moments of the model 5.52 are given by $E(Y) = XB$ and $VAR(Y) = I \otimes \Sigma$ where I is the identity matrix, Σ is the covariance matrix of the error terms and the symbol $\otimes$ denotes the kronecker product.

Each dependent variable follows a univariate model given by

$$Y_{(i)} = XB_{(i)} + \varepsilon_{(i)} \quad i = 1, 2, ..., p \tag{5.54}$$

is called the usual multiple regression model with $Cov(\varepsilon_{(i)}) = \sigma_{ii}I \equiv \sigma^2 I$. However, the errors for different responses on the same trial can be correlated.

Assumptions

To have multivariate regression model hold, we impose the following assumptions and constraints on the quantities of the model in 5.52.

- $$n \geq p + q \tag{5.55}$$

- The total number of parameters

$$m \leq pq + \frac{p(p+1)}{2} \tag{5.56}$$

- $$rank(X) = q \tag{5.57}$$

This condition is required so that we obtain a unique solution to the normal equations. If equation 5.57 is not satisfied, we use a generalized inverse.

- $$E_{n \times p} \sim N_{np}(0, \Sigma_{p \times p} \otimes I_{n \times n}) \tag{5.58}$$

In this work, we assume the error matrix E is multivariate normally distributed. The negative log likelihood of B and Σ is given by

$$-logL(B, \Sigma) = \frac{1}{2}nplog(2\pi) + \frac{n}{2}log|\Sigma| + \frac{1}{2}tr\Sigma^{-1}(Y - XB)^T(Y - XB) \tag{5.59}$$

From the equation 5.59, we can estimate the parameters B and Σ. The parameter B is estimated by

$$\hat{B} = (X^{'}X)^{-1}X'Y \tag{5.60}$$

The parameter Σ is given by

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{B})^T(Y - X\hat{B}) \tag{5.61}$$

Information Criteria

The various information criteria's for the multivariate regression case are briefly described in this section. For now, we assume that the residuals are multivariate normally distributed. Let k be the number of parameters in the model. The value of k is computed by

$$k = pq + \frac{p(p+1)}{2} \tag{5.62}$$

AIC, [Akaike, 1973], is given by

$$AIC = nlog(2\pi) + nlog|\hat{\Sigma}| + np + 2k \tag{5.63}$$

ICOMP, ([Bozdogan, 1987], [Bozdogan, 1988]) is given by

$$
\begin{aligned}
ICOMP(\hat{\beta}, (\hat{E}/\hat{B})) &= nplog(2\pi) + nlog|\hat{\Sigma}| + np + \\
&\quad (n+q)C_1(\hat{\Sigma}) + pC_1((X^{'}X)^{-1})
\end{aligned}
\tag{5.64}
$$

Future [Bozdogan and Magnus, 2003] derived $\hat{F}$, the estimated inner product form of the Fisher information matrix. It is given by

$$
\hat{F} = \begin{bmatrix} \hat{\Sigma}^{-1} \otimes X^{'}X & 0 \\ & \\ 0 & \frac{n}{2} D_p^{'}(\hat{\Sigma}^{-1} \otimes \hat{\Sigma}^{-1})D_p \end{bmatrix}
\tag{5.65}
$$

The inverse of $\hat{F}$ is given by

$$
\hat{F}^{-1} = \begin{bmatrix} \hat{\Sigma} \otimes (X^{'}X)^{-1} & 0 \\ & \\ 0 & \frac{n}{2} D_p^{+}(\hat{\Sigma} \otimes \hat{\Sigma})D_p^{+^{'}} \end{bmatrix}
\tag{5.66}
$$

The upper left block in equation 5.66 is the Kronecker product of $\hat{\Sigma}$ and $(X^{'}X)^{-1}$, which have dimensions $p \times p$ and $q \times q$ respectively, giving dimensions for the product of $pq \times pq$. The middle term of the lower right block is the Kronecker product of $\hat{\Sigma}$ and $\hat{\Sigma}$, which has dimensions $p^2 \times p^2$. The matrix D_p^{+} is the Moore-Penrose inverse of the duplication matrix D_p. The duplication matrix D_p has dimensions $p^2 \times \frac{1}{2}p(p+1)$, and so its Moore-Penrose inverse

$$
D_p^{+} = (D_p^{'}D_p)^{-1}D_p^{'}
\tag{5.67}
$$

has dimensions $\frac{1}{2}p(p+1) \times p^2$. It follows that the dimensions of the product $D_p^{+}(\hat{\Sigma} \otimes \hat{\Sigma})D_p^{+^{'}}$ are $\frac{1}{2}p(p+1) \times \frac{1}{2}p(p+1)$. This means that $s = dim(\hat{F}^{-1})$ is $pq + \frac{1}{2}p(p+1)$, which happens to be equal to the number of parameters. Further, note that the duplication matrix D_p is implicitly defined by

$$
vec(\hat{\Sigma}) = D_p vech(\hat{\Sigma})
\tag{5.68}
$$

where $vec(\hat{\Sigma})$ vectorises the distinct elements of $\hat{\Sigma}$ by vertically stacking those on and below the principal diagonal. Consequently

$$vech(\hat{\Sigma}) = D_p^+ vec(\hat{\Sigma}) \tag{5.69}$$

The outer product form of the estimated Fisher information matrix, $\hat{R}$ is given by

$$\hat{R} = \begin{bmatrix} \hat{\Sigma}^{-1} \otimes X^{'}X & \frac{1}{2}(\hat{\Sigma}^{-\frac{1}{2}} \otimes X^{'})\hat{\Gamma_1} D_p^{+^{'}} \hat{\triangle}) \\ & \\ \frac{1}{2}\hat{\triangle} D_p^+ \hat{\Gamma_1}(\hat{\Sigma}^{-\frac{1}{2}} \otimes X^{'}) & \frac{1}{4}\hat{\triangle} D_p^+ \hat{\Gamma_2^*} D_p^{+^{'}} \hat{\triangle} \end{bmatrix} \tag{5.70}$$

Here,

$$\hat{\triangle} = D_p^{'}(\hat{\Sigma}^{-1-\frac{1}{2}} \otimes \hat{\Sigma}^{-1-\frac{1}{2}})D_p \tag{5.71}$$

and

$$\hat{\Gamma_2^*} = \hat{\Gamma}_2 - n^2 I_p I_p^{'} \tag{5.72}$$

is the kurtosis matrix and $\hat{\Gamma_1}$ is the skewness matrix.

The inverse Fisher information matrix (IFIM) form of ICOMP is given by

$$\begin{aligned} ICOMP(IFIM) &= nplog(2\pi) + nlog|\hat{\Sigma}| + np + C_1(\hat{F^{-1}}) \qquad (5.73) \\ &= nplog(2\pi) + nlog|\hat{\Sigma}| + np \\ &\quad + \frac{p(p+q)}{2}A - \frac{1}{2}(p+q+1)log|\hat{\Sigma}| \\ &\quad - \frac{p}{2}log|(X^{'}X)^{-1}| - \frac{p}{2}log(2) \end{aligned}$$

where

$$A = log[\frac{tr(\hat{\Sigma})tr(X^{'}X)^{-1} + \frac{1}{2}tr(\hat{\Sigma}^2) + \frac{1}{2}(tr(\hat{\Sigma}))^2 + \sum_j \hat{\sigma_{jj}^2}}{p(p+q)}]$$

The mis-specification form of ICOMP is given by

$$ICOMP(IFIM)_{MISSPEC} = nplog(2\pi) + nlog|\hat{\Sigma}| + np + 2C_1(\hat{F^{-1}}\hat{R}\hat{F^{-1}}) \tag{5.74}$$

Algorithm: Optimal Scaling Method

This algorithm fits a multivariate regression for a set of continuous responses, Y and a mixed set of predictors, X, and also selects the optimal predictors that explain most of the variation in the response set, Y.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Continuous Response Data set: Y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Run the Gifi transformation on the data set, X, and optimally scale the categorical variables in the Gifi space.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model in the Gifi space. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}.
    - Perform multivariate regression with Y as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

   - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.

   - Go to step 3

Algorithm: Linear Combination Method

This algorithm fits a multivariate regression for a set of continuous responses, Y and a mixed set of predictors, X, and also selects the optimal predictors that explain most of the variation in the response set, Y.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Continuous Response Data set: Y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Transform the mixed data set X to a pure continuous space using Gifi transformation.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

   - Build a new predictor data matrix, X_{new}. Since X_{new} might be a mixed data set, we split the X_{new} matrix into X_{con} and X_{cat} where X_{con} is the data on

the continuous predictors and X_{cat} is a 1-dimensional continuous data of the categorical predictors in the Gifi space. Hence X_{new} can be represented as $X_{new} = \left[\begin{array}{cc} X_{con} & X_{cat} \end{array}\right]$.

- Perform multivariate regression with Y as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N-1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

   - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
   - Go to step 2

5.2.5 Multivariate Logistic Regression

Methodology

Let a data set, D_0 consists of 2 categorical dependent variables, y_1, y_2 and 5 independent variables, $x_1, x_2, ..., x_5$ where x_1 and x_4 are continuous and x_2, x_3 and x_5 are categorical. The data set, D_0 is a perfect example of a multivariate logistic regression setting. In the Gifi space, the categorical dependent variables are transformed to a 1-dimensional continuous space on the response side and the categorical independent variables are transformed to a 1-dimensional continuous space on the predictors side. Suppose, if y_1 contains 2 categories and y_2 contains 2 categories, a linear combination of the categories of the two dependent variables y_1 and y_2 is transformed to the continuous space with $2 \times 2 = 4$ unique continuous values. Since we can treat each value as a unique class, this becomes a multi class logistic regression in the Gifi space. If y_1 and y_2 contains many categories, then their linear combination in the Gifi space will contain many unique continuous values. In this case, it would make reasonable sense to treat the problem as a multiple regression problem.

Algorithm: Optimal Scaling Method

This algorithm fits a multivariate logistic regression for a categorical response set, Y, and a mixed set of predictors, X, and also selects the optimal predictors that explain most of the variation in the response.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Categorical Response Data set : Y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Run the Gifi transformation on the data set, X and Y, and optimally scale the categorical variables in the Gifi space. The response Y in the Gifi space becomes continuous. Hence, we can fit a multivariate regression with Y in the Gifi space as response and X in the Gifi space as predictors.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}, in the Gifi space.
    - Run multivariate regression with Y in the Gifi space as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

    - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
    - Go to step 4

Algorithm: Linear Combination Method

This algorithm fits a multivariate logistic regression for a categorical response set, Y, and a mixed set of predictors, X, and also selects the optimal predictors that explain most of the variation in the response.

Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Categorical Response Data set : Y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Transform the mixed data set X to a pure continuous space using Gifi transformation.
2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.
3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}. Since X_{new} might be a mixed data set, we split the X_{new} matrix into X_{con} and X_{cat} where X_{con} is the data on

the continuous predictors and X_{cat} is a 1-dimensional continuous data of the categorical predictors in the Gifi space. Hence X_{new} can be represented as $X_{new} = \left[\begin{array}{cc} X_{con} & X_{cat} \end{array} \right]$.

- Since the data set, Y is a categorical data set, transform that to a 1-dimensional continuous space, called Y_{new}, by the Gifi transformation. Suppose $Y = \{y_1, y_2\}$. Let y_1 and y_2 contains two categories each. Therefore, the 1-dimensional continuous data (linear combination of the categories of y_1, y_2 in the Gifi space) would contain $2 \times 2 = 4$ distinct continuous values. We can consider these 4 distinct continuous values as 4 distinct classes.

- Perform multi-class logistic regression with Y_{new} as response and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

   - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.

   - Go to step 2

5.3 Discriminant Analysis

5.3.1 Introduction

Discriminant Analysis (DA) is a supervised classification technique. DA consists of assigning or classifying an individual or object to one of several known or unknown alternative classes (or groups) on the basis of many measurements on the individuals or objects, or cases, [Bozdogan, 2005]. The goal of discriminant analysis is: given the data set with two or more than two classes (or groups), say, what is the best feature or feature set either linear or non-linear to discriminate between the classes and maximize average class separation. It uses the density estimation strategy and assume a parametric model for the densities, [Wasserman, 2004]. We assume that the set of predictors are multivariate gaussian distributed.

5.3.2 Linear and Quadratic Discriminant Analysis

A discrimination procedure can be developed via the estimation of the class conditional density functions and the use of Bayes rule. We consider the matrix consisting of n total observations on p variables on $k = 1, 2..., K$ classes or groups given by

$$X = \begin{bmatrix} x_{11}^{'} \\ \vdots \\ x_{1n_1}^{'} \\ x_{21}^{'} \\ \vdots \\ x_{2n_1}^{'} \\ \vdots \\ x_{kn_1}^{'} \\ \vdots \\ x_{kn_k}^{'} \end{bmatrix}_{(n_1+...+n_k)\times p}$$

Let $f_k(x)$ denote the class-conditional density of X in k-th class with prior probability π_k of class k such that $\sum_{k=1}^{K} \pi_k = 1$. We classify an observation to a class for which the posterior probability of group membership is the greatest. This is achieved by utilizing the Bayes rule or theorem

$$P(K = k|X = x) = \frac{f_k(x)\pi_k}{f(x)} = \frac{f_k(x)\pi_k}{\sum_{k=1}^{K} f_k(x)\pi_k} \tag{5.75}$$

Applying the Bayes rule in equation 5.75 and by considering the class conditional density to be a multivariate gaussian model given by

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^{'}\Sigma_k^{-1}(x-\mu_k)}, \tag{5.76}$$

we obtain the log posterior probability of group membership as

$$\begin{aligned} logP(K = k|X = x) &= log(f_k(x)) + log(\pi_k) - log(f(x)) \\ &= -\frac{1}{2}(x - \mu_k)^{'}\Sigma_k^{-1}(x - \mu_k) - \frac{1}{2}log|\Sigma_k| \\ &\quad -\frac{p}{2}log(2\pi_k) - log(f(x)) \end{aligned} \tag{5.77}$$

In equation 5.77 since log$f(x)$ is independent of the class, the discrimination rule can be established. So in comparing two classes k and l, we assign the observation vector x to class k if

$$d_k(x) > d_l(x) \tag{5.78}$$

for all $k \neq l$, where

$$d_k(x) = log(\pi_k) - \frac{1}{2}log|\Sigma_k| - \frac{1}{2}(x - \mu_k)^{'}\Sigma_k^{-1}(x - \mu_k) \tag{5.79}$$

Classifying an observation vector x on the basis of the values of $d_k(x), k = 1, 2, ..., K$ is called the gaussian-based quadratic discriminant function (QDA), [McLachlan, 1992], since the decision boundary between each pair of classes k and l is described by a quadratic equation in x.

We note that in the special case when the class covariance matrices $\Sigma_1, ..., \Sigma_K$ are all the same, that is, $\Sigma_k = \Sigma$ for all k, the linear discriminant analysis (LDA) arises and that from equation 5.79 we obtain the linear discriminant functions given by

$$d_k(x) = log(\pi_k) - \frac{1}{2}\mu_k^{'}\Sigma^{-1}\mu_k + x^{'}\Sigma^{-1}\mu_k. \tag{5.80}$$

In comparing two classes k and l, we assign the observation vector x to class k if

$$d_k(x) > d_l(x) \tag{5.81}$$

for all $k \neq l$, where $d_k(x)$ is given in equation 5.80.

In practice we do not know the parameters of the Gaussian model. We need to estimate these unknown parameters by using the maximum likelihood method. Hence, the estimation problem basically reduces to the parameter estimation for the class conditional densities. Given that the classes are known for the training data, we can write down the likelihood or the log likelihood of the (MLE's) or plug-in-estimators given below.

- The estimator of the prior probability π_k is $\hat{\pi_k} = \frac{n_k}{n}$, where n_k is the number of observations in class k.
- $\hat{\mu_k} = \bar{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}$, are the sample means of each class; $\bar{x}_k = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k} x_{ki}$ is the sample mean or grand mean of the entire training set, and $n = \sum_{k=1}^{K} n_k$ is the total number of observations, $k = 1, 2, ..., K$.
- $\hat{\Sigma_k} = S_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^{'}$ is the estimated class covariance or scatter matrix for $k = 1, 2, ...K$.
- Averaged over all classes the biased scatter matrix describing the noise is:

  $$\hat{\Sigma} = S_W = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^{'} \tag{5.82}$$

  This matrix is called the within-scatter (or inter-scatter) matrix. It describes the average scattering within classes. An unbiased version of S_W is given by

  $$\hat{\Sigma} = (\frac{n}{n-K})S_W \tag{5.83}$$

- Between-scatter (or intra-scatter) matrix S_B that describes the scattering of the class-dependent sample means around the overall average is given by

$$S_B = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^{'} \tag{5.84}$$

- Total scatter matrix is:

$$\begin{aligned} S_T &= S_W + S_B \\ &= \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^{'}. \end{aligned} \tag{5.85}$$

Both the LDA and QDA are computationally efficient. LDA is expected to perform well in homoscedastic cases when sample size is large compared to the dimension of the measurement or variable space. In a well-determined homoscedastic case, LDA should give better performance because it estimates fewer parameters.

On the other hand, QDA should perform well in a well-determined heteroscedastic case. However, LDA may fail when the within class distributions are heteroscedastic. Also, both LDA and QDA will have problems when any of the matrices $\hat{\Sigma_k}$ is singular. For both LDA and QDA will have problems when the data are nonlinear and when the class conditional densities do not follow a Gaussian distribution. Since the data is continuous in the Gifi space, we overcome the problem that occurs when the data is nonlinear.

5.3.3 Information Criteria

In this section, we briefly list the information criteria measures for choosing the number of discriminant functions. The number of useful discriminant functions is defined by the number of nonzero eigenvalues in the classes, which in turn, equal to the rank of S_B, i.e.,

$$m = Number of DF's = rank(S_B), m = 0, 1, 2, ..., s. \tag{5.86}$$

Let $f_k(x, \theta)$ denote the class conditional density of X, where the parameter vector θ is:

$$\theta = (\mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K). \tag{5.87}$$

Then -2 times maximized log-likelihood function is

$$\begin{aligned} -2l(\hat{\mu_k}, \hat{\Sigma}) &= -2logL(\hat{\mu_k}, \hat{\Sigma}) \\ &= nplog(2\pi) + np + nlog|S_W| + \\ & \quad nlog \prod_{i=m+1}^{s} (1 + \lambda_i) \end{aligned} \tag{5.88}$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m \geq \lambda_{m+1} = ... = \lambda_p = 0$ are the eigenvalues of $S_W^{-1} S_B$ and $\lambda_1, \lambda_2 ..., \lambda_m$ are the positive eigenvalues.

The selection based on AIC is equivalent to the procedure by which we choose the DF $d_k(x)$ such that the criterion differences of AIC

$$DIC(m) = nlog \prod_{i=m+1}^{s} (1 + \lambda_i) - 2(p - m)(K - m) \tag{5.89}$$

is minimum, where n is the total number of observations, p is equal to the number of variables, and K is the number of classes.

Similarly, we can compute $DICOMP_{1F}(m)$ given by

$$\begin{aligned} DICOMP_{1F}(m) &= nlog \prod_{i=m+1}^{s} (1 + \lambda_i) - 2C_{1F}(S_W^{-1} S_B) \\ &= nlog \prod_{i=m+1}^{s} (1 + \lambda_i) - 2[\frac{1}{4\bar{\lambda}_a^2} \sum_{j=1}^{p} (\lambda_j - \bar{\lambda}_a)^2] \end{aligned} \tag{5.90}$$

where $\bar{\lambda}_a = \frac{1}{m} \sum_{j=1}^{m} \lambda_j$ is the arithmetic mean of the eigenvalues, and $\bar{\lambda}_g = (\prod_{j=1}^{m} \lambda_j)^{1/m}$ is the geometric mean of the eigenvalues of $S_W^{-1} S_B$, respectively. We note that $C_{1F}(.) \geq 0$ with $C_{1F}(.) = 0$ only when all $\lambda_j = \bar{\lambda}$. Also, $C_{1F}(.)$ measures the relative variation in the eigenvalues while $C_F(.)$ measures the absolute variation in the eigenvalues.

5.3.4 Algorithm: Optimal Scaling Method

This algorithm performs a supervised classification (discriminant analysis) for a categorical response and a mixed set of predictors, X, and also selects the optimal predictors that classifies the data into k classes where k is the number of categories of the response variable.
Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Categorical Response Data : y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Run the Gifi transformation on the data set, X, and optimally scale the categorical variables in the Gifi space.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}.
    - Perform discriminant analysis with y as classification variable and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

- Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
- Go to step 3

5.3.5 Algorithm: Linear Combination Method

This algorithm performs a supervised classification (discriminant analysis) for a categorical response and a mixed set of predictors, X, and also selects the optimal predictors that classifies the data into k classes where k is the number of categories of the response variable.
Input:

Maximum Iteration : maxIter
Probability of Cross over : pCrossover
Probability of Mutation : pMutation
Cross over type : Uniform, Single point, Two point
Population size : N
Predictor Data : X
Categorical Response Data : y
Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Transform the mixed data set X to a pure continuous space using Gifi transformation.
2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.
3. For each chromosome in the population
    - Build a new predictor data matrix, X_{new}. Since X_{new} might be a mixed data set, we split the X_{new} matrix into X_{con} and X_{cat} where X_{con} is the data on the continuous predictors and X_{cat} is a 1-dimensional continuous data of the categorical predictors in the Gifi space. Hence X_{new} can be represented as $X_{new} = \left[\begin{array}{cc} X_{con} & X_{cat} \end{array} \right]$.

- Perform discriminant analysis with y as classification variable and X_{new} as predictors and compute the respective information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N-1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

   - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
   - Go to step 2

5.4 Cluster Analysis

5.4.1 Introduction

Clustering is the classification of objects into different groups. It partitions a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait. There are two different types of clustering algorithms namely - hierarchical and partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once. Hierarchical algorithms can be agglomerative (”bottom-up”) or divisive (”top-down”). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. There are several clustering algorithms such as k-means, fuzzy clustering, medoids etc. In this work, we do an unsupervised clustering of the data in the Gifi space assuming that the data is generated from a mixture of gaussian distributions.

5.4.2 Gaussian Mixtures

Model

The problem of clustering of n individuals on the basis of p-dimensional observation vectors $x_1, x_2, ..., x_n \in \Re^p$ will be studied using a mixture of normal probability density functions [Bozdogan, 1994]. In this method, we do not know a priori the number of clusters (K), mixing proportions, mean vectors, and covariance matrices of the class distributions. If we assume that each observation vector x_i has probability π_k of coming from the k-th population $k \in 1, 2, ..., K$, then $x_1, x_2, ..., x_n$ is a sample from

$$f(x) \equiv f(x; \pi, \mu, \Sigma) = \sum_{k=1}^{K} \pi_k g_k(x; \mu_k, \Sigma_k), \tag{5.91}$$

where $\pi = (\pi_1, \pi_2, ..., \pi_{K-1})$ are K-1 independent mixing proportions such that

$$0 \leq \pi_k \leq 1 \qquad for \quad k = 1, 2, ..., K \qquad and \qquad \pi_K = 1 - \textstyle\sum_{k=1}^{K-1} \pi_k \tag{5.92}$$

and where $g_k(x; \mu_k, \Sigma_k)$ is the k-th component multivariate normal density function given by

$$g_k(x; \mu_k, \Sigma_k) = (2\pi)^{-p/2} |\Sigma_k|^{-1/2} e^{-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)}. \tag{5.93}$$

The model given in equation 5.93 is called the standard multivariate normal mixture model.

Parameter Estimation

In this mixture, the parameters to be estimated are K-1 mixing proportion estimates, K mean vectors, K covariance matrices. We can write the loglikelihood function of the data $x_1, x_2, ..., x_n$ as:

$$\begin{aligned} l(\theta) &\equiv logL(\theta|X) \\ &= \sum_{i=1}^{n} log[\sum_{k=1}^{K} \pi_k (2\pi)^{-p/2} |\Sigma_k|^{-1/2} e^{-\frac{1}{2}(x_i-\mu_k)'\Sigma_k^{-1}(x_i-\mu_k)}] \qquad (5.94) \\ & \qquad (5.95) \end{aligned}$$

To obtain the maximum likelihood estimators (MLE's) of the unknown parameters, we use matrix differential calculus and compute the partial derivatives of the log likelihood function $l(\theta)$ with respect to π_k, the mean vector μ_k, and Σ_k, respectively, and set these equal to zero. We obtain the following ML equations:

$$P(\hat{k}|x_i) = \frac{\hat{\pi}_k g_k(x_i; \hat{\mu_k}, \hat{\Sigma_k})}{\sum_{k=1}^{K} \hat{\pi}_k g_k(x_i; \hat{\mu_k}, \hat{\Sigma_k})} \qquad k = 1, 2, ..., K \tag{5.96}$$

$$\hat{\pi_k} = \frac{1}{n} \sum_{i=1}^{n} P(\hat{k}|x_i) \qquad k = 1, 2, ..., K \tag{5.97}$$

$$\hat{\mu_k} = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} x_i P(\hat{k}|x_i) \qquad k = 1, 2, ..., K \tag{5.98}$$

$$\hat{\Sigma}_k = \frac{1}{n\hat{\pi}_k} \sum_{i=1}^{n} \hat{P}(k|x_i)(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)' \qquad k = 1, 2, ..., K \tag{5.99}$$

where $\hat{\pi}_k$ is the estimated mixing proportion π_k, $\hat{\mu}_k$ is the estimated mean vector μ_k, $\hat{\Sigma}_k$ is the estimated covariance matrix Σ_k and $\hat{P}(k|x_i)$ is the estimated posterior probability of group membership of the observation vector x_i in the cluster k.

Information Criteria

In this section, we briefly state the information criteria such as AIC, CAIC and ICOMP for the gaussian mixture case. Let p be the dimension of a model M. Let $L(\hat{\theta})$ be the maximum likelihood value of the model M. Let K be the number of mixtures fitted to the data. AIC, [Akaike, 1973], for the model M is given by

$$AIC = -2log(L(\hat{\theta})) + 3m \tag{5.100}$$

where m is the number of free parameters estimated within the model M. There are K-1 mixture parameters, Kp mean parameters and $Kp(p + 1)/2$ covariance parameters to estimate. Hence m is given by

$$m = K - 1 + K \times p + K \times p \times (p + 1)/2 \tag{5.101}$$

CAIC, [Bozdogan, 1987], for the model M is given by

$$CAIC = -2log(L(\hat{\theta})) + m(log(n) + 1) \tag{5.102}$$

where n is the number of observations in the data.

ICOMP, ([Bozdogan, 1987], [Bozdogan, 1988], [Bozdogan, 90a], [Bozdogan, 90b], [Bozdogan, 2004], [Bozdogan, 2005]), is given by

$$ICOMP = -2log(L(\hat{\theta})) + 2C_1(\hat{F}^{-1}) \tag{5.103}$$

where $\hat{F}^{-1}$ is the estimated Inverse Fisher Information Matrix (IFIM) and $C_1(\hat{F}^{-1})$ is given by

$$C_1(\hat{F}^{-1}) = \frac{s}{2}log[\frac{trace(\hat{F}^{-1})}{s}] - \frac{1}{2}log|\hat{F}^{-1}| \tag{5.104}$$

The parameter s in the equation 5.104 is the rank($\hat{F}^{-1}$).

Let $\lambda_1, \lambda_2, ..., \lambda_s$ be the eigenvalues of $\hat{F}^{-1}$. Further let $\bar{\lambda}_a = \frac{1}{s}\sum_{j=1}^{s}\lambda_j$ be the arithmetic mean and $\bar{\lambda}_g = (\prod_{j=1}^{s}\lambda_j)^{1/s}$ is the geometric mean of the eigenvalues. The complexity of $\hat{F}^{-1}$ can be written as $C_1(\hat{F}^{-1}) = \frac{s}{2}log(\frac{\bar{\lambda}_a}{\bar{\lambda}_g})$. Hence, ICOMP can be given by

$$ICOMP = -2log(L(\hat{\theta})) + slog(\frac{\bar{\lambda}_a}{\bar{\lambda}_g}) \tag{5.105}$$

The $\hat{F}^{-1}$ for the model M [Bozdogan, 1994] is given by

$$\hat{F}^{-1} = Diag(\hat{F_1}^{-1}, ..., \hat{F_K}^{-1}) \tag{5.106}$$

where $\hat{F_k}^{-1}$ is given by

$$\hat{F_k}^{-1} = \begin{bmatrix} \frac{1}{\hat{\pi_k}\hat{\Sigma_k}} & 0 \\ 0 & 2D_p^{+}(\hat{\Sigma_k} \otimes \hat{\Sigma_k})D_p^{+'} \end{bmatrix} \tag{5.107}$$

$\hat{F}^{-1}$ in equation 5.106 is due to the fact that the parameters for a particular mixture cluster are independent of the parameters of the subsequent mixture cluster after the clusters are recovered. Hence, $\hat{F}^{-1}$ is a block diagonal matrix with the diagonal block given by the estimated asymptotic covariance matrices $\hat{F_k}^{-1}$ for the k-th mixture cluster, and $\otimes$ denotes the Kronecker product. In equation 5.107, $D_p^+ = (D_p^{'}D_p)^{-1}D_p^{'}$ is the Moore-Penrose inverse of the duplication matrix D_p. A duplication matrix is a unique $p^2p(p+1)/2$ matrix which transforms v$(\hat{\Sigma_k})$ into vec$(\hat{\Sigma_k})$. v$(\hat{\Sigma_k})$ denotes the $p(p+1)/2$-vector that is obtained from vec$(\hat{\Sigma_k})$ by eliminating all supra-diagonal elements of $\hat{\Sigma_k}$ and stacking the remaining columns one underneath the other.

For computational efficiency, we expand the equation 5.103 for model M as

$$ICOMP = -2\sum_{i=1}^{n} log[\sum_{k=1}^{K} \pi_k g_k(x;\mu_k,\Sigma_k)] + [kp + kp(p+1)/2] \times C \tag{5.108}$$

where C is given by

$$C = log[\frac{\sum_{k=1}^{K}\frac{1}{\hat{\pi_k} trace(\hat{\Sigma_k})} + \frac{1}{2}trace(\hat{\Sigma_k}^{2}) + \frac{1}{2}(trace(\hat{\Sigma_k}))^2 + \sum_{j=1}^{p}(\hat{\sigma}_{kjj})^2}{kp + kp(p+1)/2}]$$

Algorithm: Optimal Scaling Method

This algorithm performs an unsupervised classification of a mixed data set, X, and also selects the optimal predictors that classifies the data into the target number of groups. We first, perform an unsupervised clustering (assuming mixtures of gaussian distribution) on X to determine the optimal number of gaussian mixtures, $optMix$. We now perform a variable selection method to select the optimal predictors that classifies the data into $optMix$ groups.

Input:

Maximum Iteration : maxIter

Probability of Cross over : pCrossover

Probability of Mutation : pMutation

Cross over type : Uniform, Single point, Two point

Population size : N

Data set : X

Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Run the Gifi transformation on the data set, X, and optimally scale the categorical variables in the Gifi space.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}.
    - Perform cluster analysis with X_{new} and number of mixtures, $optMix$. Also compute the information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

    - Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
    - Go to step 3

Algorithm: Linear Combination Method

This algorithm performs an unsupervised classification of a mixed data set, X, and also selects the optimal predictors that classifies the data into the target number of groups. We first, perform an unsupervised clustering (assuming mixtures of gaussian distribution) on X to determine the optimal number of gaussian mixtures, $optMix$. We now perform a variable selection method to select the optimal predictors that classifies the data into $optMix$ groups.

Input:

Maximum Iteration : maxIter

Probability of Cross over : pCrossover

Probability of Mutation : pMutation

Cross over type : Uniform, Single point, Two point

Population size : N

Data set : X

Information score : AIC, ICOMP, ICOMPIFIM, CAIC, SBC

1. Transform the mixed data set X to a pure continuous space using Gifi transformation.

2. Generate a random population of size N and dimension p, where p is the number of predictors in the model. Consider each row of the population to be a chromosome.

3. For each chromosome in the population

    - Build a new predictor data matrix, X_{new}. Since X_{new} might be a mixed data set, we split the X_{new} matrix into X_{con} and X_{cat} where X_{con} is the data on the continuous predictors and X_{cat} is a 1-dimensional continuous data of the categorical predictors in the Gifi space. Hence X_{new} can be represented as $X_{new} = \begin{bmatrix} X_{con} & X_{cat} \end{bmatrix}$.
    - Perform cluster analysis with X_{new} and number of mixtures, $optMix$. Also compute the information score.

4. Sort the chromosome in the population in the increasing order of the information score. The chromosome with the lowest information score is considered to be the best chromosome than the $N - 1$ other chromosomes.

5. Stop if the stopping criteria is met and return the best model from the current population or else

- Perform cross over and mutation with pCrossover, pMutation and the cross over type to generate a new population. Always include the best model in the new population.
- Go to step 2

Chapter 6

Numerical Results

6.1 Mixtures of Multivariate Bernoulli Distributed Data

We ran our experiments on one simulated and two real data sets to illustrate the identification process of a number of mixtures in a Multivariate Bernoulli distributed data using genetic algorithm and information complexity.

6.1.1 A simulated data example

First, we show the results on the simulated data set. The data is simulated as described below.

1. Simulate random normals of 50 rows and 5 columns from N(0, 1).

2. Simulate random normals of 50 rows and 5 columns from N(2, 1).

3. Create a matrix of 100 rows and 5 columns where the first 50 rows of the matrix are from step 1 and the second 50 rows of the matrix are from step 2.

4. Convert the generated random normals to binary form by the following criteria. If each element of the generated data matrix is greater than the mean of the random normals in that column assign a value 1 or else assign a value 0.

5. Run the Bernoulli Mixture algorithm on the generated binary data set.

Table 6.1: Simulated Data: Lack of fit & information criteria values for mixtures 1 to 6

# of Mixtures	*LackofFit*	*AIC*	*SBC*	*ICOMP*	*ICOMP (PEU)*
1	691.7854	706.7854	714.8113	706.7855	706.7855
2	536.7557	**569.7557**	**587.4126**	**575.4775**	**582.9306**
3	529.8647	580.8647	608.1526	587.081	595.1783
4	527.9289	596.9289	633.8478	603.42	611.8751
5	527.6492	614.6492	661.1991	621.1651	629.6527
6	527.1351	632.1351	688.3161	639.0032	647.9495
7	523.3198	646.3198	712.1318	653.5446	662.9554
8	516.55	657.55	732.993	664.9358	674.5563
9	**516.5408**	675.5408	760.6148	682.9077	692.5037
10	516.5457	693.5457	788.2507	700.9432	710.7542

Table 6.2: Simulated Data: Count of the number of mixtures by Lack of fit & ICOMP

# of Mixtures	*Count of Mixtures selected by LackofFit*	*Count of Mixtures selected by ICOMP*
1	0	0
2	0	100
3	12	0
4	2	0
5	4	0
6	3	0
7	10	0
8	10	0
9	23	0
10	36	0

We run 100 simulations of our multivariate Bernoulli mixture algorithm on random normals generated from two different mixtures. We list the information criteria scores for one of the simulations. All the information criteria values in table 6.1 are minimum for the two mixture model which is true since the data is generated from two mixtures of normal distribution where as the minimum lack of fit criteria chooses nine mixtures. Table 6.2 shows that out of 100 simulations, the ICOMP approach picked up the right model (two mixture case) in all cases whereas the lack of fit criteria overfitted.

6.1.2 Mobile phone data set

This data set contains information on 20 variables and contains 1021 observations. PCA is obtained on this data set and the PCA scores (20 PC scores) are used for determining the number of mixtures in this mobile phone data. A neat and straight-forward method of converting the PCA score to binary data is to assign a value 1 for all scores greater than the mean in that column and a value 0 for all scores less than the mean in that column.

Table 6.3: Mobile Phone Data: Lack of fit & information criteria values for mixtures 1 to 10

# of Mixtures	*Lackoffit*	*AIC*	*SBC*	*ICOMP*	*ICOMP (PEU)*
1	28101.3948	28161.3948	28239.9656	28161.398	28161.4056
2	28013.3025	28136.3025	28297.3726	28140.6627	28151.4073
3	27924.6817	28110.6817	28354.251	28118.6299	28138.2166
4	27801.4749	28050.4749	28376.5435	28060.3503	28084.686
5	27730.165	28042.165	28450.7329	28054.1576	28083.7106
6	27672.1705	28047.1705	28538.2377	28061.6492	28097.3287
7	27538.5697	27976.5697	28550.1362	27993.5311	28035.3288
8	27455.5341	27956.5341	28612.6	27974.8435	28019.9626
9	**27385.0512**	**27949.0512**	**28687.6163**	**27970.3776**	**28022.9315**
10	27410.9423	28037.9423	28859.0067	28057.5909	28106.0104

Table 6.4: Mobile Phone Data: Mixing proportion estimates

Mixture	*Estimate*
$\hat{\pi}_1$	0.6521
$\hat{\pi}_2$	0.0345
$\hat{\pi}_3$	0.0281
$\hat{\pi}_4$	0.0184
$\hat{\pi}_5$	0.0930
$\hat{\pi}_6$	0.0287
$\hat{\pi}_7$	0.0422
$\hat{\pi}_8$	0.0775
$\hat{\pi}_9$	0.0253

The following results (table 6.3) are generated by MATLAB for this data set. In this case, the nine mixture model has the minimum information criteria value for AIC, SBC, ICOMP and $ICOMP_{PEU}$. Even according to the maximum likelihood criteria, the nine mixture model is considered to be good since it has the minimum lack of fit.

The mixing proportion estimates and the probability estimates for each variable in each mixture in the mobile phone data set is given below in table 6.4 and table 6.6 respectively. The classification table is given in table 6.5.

We now do a variable selection using GA on this data set. The following parameters are used as inputs to the GA process: maximum iterations - 10, population size - 50, probability of crossover - 0.90, probability of mutation - 0.10 and crossover type - uniform. The GA selected the first two principal components as the optimal predictors for clustering. The associated ICOMP score for the first two principal components is 2793.3999. The plot of the best ICOMP at the end of each iteration of the GA process is shown in figure 6.1.

Table 6.5: Mobile Phone Data: Classification matrix

Mixture	*# of observations*
1	671
2	37
3	32
4	21
5	83
6	33
7	44
8	73
9	27

Table 6.6: Mobile Phone Data: Probability estimates of the variables

Probability	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
p_1	0.2861	0.1513	0.2895	0.8092	0.9881	0.2516	0.7716	0.1671	0.0366
p_2	0.4954	0.1496	0.0000	0.0000	0.8546	0.4508	0.5443	0.0006	0.0000
p_3	0.4330	0.4327	0.5055	0.3142	0.6294	0.2270	0.5514	0.3527	0.6541
p_4	0.4804	0.1466	0.0000	0.1702	0.5378	0.6530	0.5946	0.3418	1.0000
p_5	0.5577	0.3637	0.1436	0.4082	0.2767	0.5628	1.0000	0.7605	0.0000
p_6	0.4348	0.3398	0.7468	0.3514	0.3986	0.2909	0.7198	0.6724	1.0000
p_7	0.4995	0.7136	0.5420	0.1565	0.2543	0.9506	0.8797	0.2407	0.4544
p_8	0.4890	0.2989	0.6218	0.6706	0.7617	0.4534	0.4550	0.4422	0.7021
p_9	0.6171	0.0000	0.8192	0.0000	0.3999	0.1637	0.2039	0.0000	0
p_{10}	0.5744	0.8279	1.0000	0.3200	0.3231	0.4619	0.4429	0.2677	0.3537
p_{11}	0.5228	0.4674	0.2284	0.1978	0.1932	0.1895	0.3962	0.1623	0.8234
p_{12}	0.5240	1.0000	0.3321	1.0000	0.7102	1.0000	0.0565	0.3402	0.5576
p_{13}	0.5766	0.0000	1.0000	0.0000	0.5166	0.0000	0.0792	0.5060	1.0000
p_4	0.5501	0.8021	0.5612	0.4000	0.5585	0.0000	0.9962	0.2374	0.5045
p_{15}	0.3914	0.1374	1.0000	0.8063	0.5485	0.3444	0.6247	0.6130	0.9673
p_{16}	0.4193	0.5930	0.2079	0.2835	0.5200	0.0000	0.6485	0.4990	0.3019
p_{17}	0.5296	0.0000	0.8199	0.6790	0.4889	0.1624	0.8438	0.4811	0.0931
p_{18}	0.5221	0.1187	0.0000	0.5325	0.4647	0.9999	0.3422	0.5226	0.2553
p_{19}	0.4922	0.5899	0.6700	0.0000	0.4516	0.6806	0.9505	0.5453	0.1267
p_{20}	0.4062	0.2704	0.7449	0.0421	0.6057	0.5263	0.4235	0.4123	0.0000

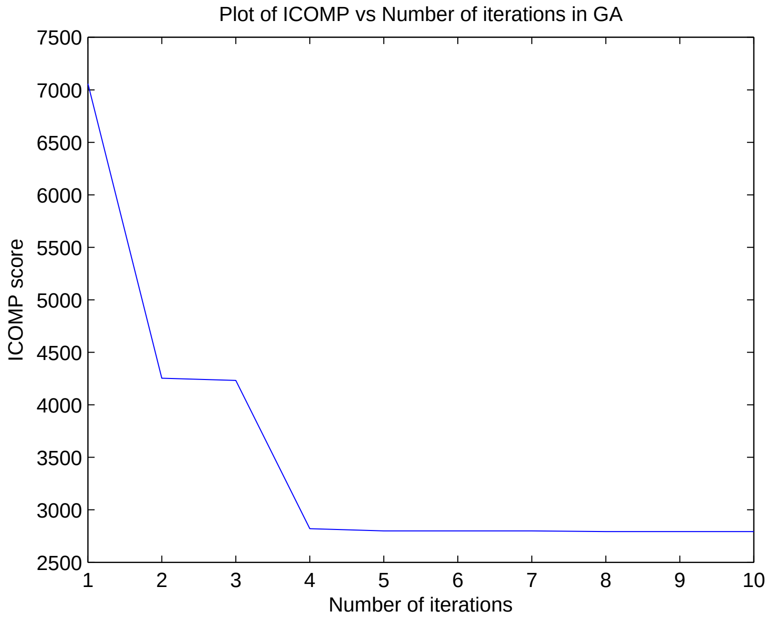


Figure 6.1: Mobile Phone Data: Plot of ICOMP vs Number of iterations in GA

6.1.3 Keratoconjunctivitissicca (KCS) data set

Now we present the results on a real medical data set given in [Botev, 2005]. The data is regarding the diagnosis of *Keratoconjunctivitissicca* (KCS). The description of the data set is briefly given below:

1. Part 1 consists of 40 patients suffering from KCS. Each patient may or may not have any of the 10 possible symptoms of the disease. The presence of the symptoms is represented as binary row vectors of length 10. A 1 means that the symptom is present and a 0 stands for no clinically obvious pathology.

2. Part 2 consists of 37 non-KCS patients.

3. Part 1 and Part 2 form the first group of 77 patients, referred to as group-1.

4. The same 10 symptoms are recorded for another group of 41 patients, henceforth referred to as the group-2 patients. This group consists of 24 KCS patients and 17 non KCS patients.

Table 6.7: KCS Group1: Lack of fit & information criteria values for mixtures 1 to 10

# of	*Lackoffit*	*AIC*	*SBC*	*ICOMP*	*ICOMP*
Mixtures					*(PEU)*
1	888.2047	918.2047	931.6428	918.5119	918.8718
2	715.5283	**778.5283**	**806.7482**	**793.3299**	**810.6759**
3	702.5933	798.5933	841.5951	814.974	834.1706
4	696.9249	825.9249	883.7086	842.9794	862.9655
5	696.8369	858.8369	931.4024	875.9241	895.9487
6	693.4241	888.4241	975.7715	905.5905	925.7079
7	694.4588	922.4588	1024.588	939.4061	959.2667
8	693.161	954.161	1071.0721	971.5217	991.8667
9	691.9613	985.9613	1117.6543	1003.7215	1024.5348
10	**690.655**	1017.655	1164.1298	1035.6503	1056.7391

Table 6.8: KCS Group1: Mixing proportion estimates

Mixture	*Estimate*
$\hat{\pi_1}$	0.4764
$\hat{\pi_2}$	0.5236

Running the Multivariate Bernoulli algorithm on the group 1 data set generated the following information criteria values shown in table 6.7. Even in this case, the two mixture model has the minimum information criteria value for AIC, SBC, ICOMP, $ICOMP_{PEU}$. According to the maximum likelihood criteria, the six mixture model is considered to be good since it has the minimum lack of fit.

The estimated parameters for the two mixture model in the group 1 data set are given below in table 6.8 and table 6.10 respectively. Table 6.9 shows that out of 77 observations, the two mixture model classified 37 observations into mixture 1 and 40 observations into mixture 2. The confusion matrix for the above classification is given in table 6.11. We now present the results of our GA algorithm to determine which subset of the ten symptoms in the group 1 data set are sufficient for classification into the target number of mixtures. The probability of mutation is 0.10. The population size is taken to be 50. The GA is run for 10 iterations. The following results (table 6.12) are generated at the end of

Table 6.9: KCS Group1: Classification matrix

Mixture	*# of observations*
1	37
2	40

Table 6.10: KCS Group1: Probability estimates of the variables

Probability	$m = 1$	$m = 2$
p_1	0.8364	0.0823
p_2	0.8315	0.0372
p_3	0.7321	0.0283
p_4	0.7363	0.0493
p_5	0.4902	0.0749
p_6	0.2727	0.0247
p_7	0.4359	0.2483
p_8	0.4085	0.0252
p_9	0.2467	0.0484
p_{10}	0.4090	0.0495

Table 6.11: KCS Group1: Confusion matrix

$Actual \backslash Predicted$	$Mixture1$	$Mixture2$
$Mixture1$	36	4
$Mixture2$	1	36

each iteration. The results of our GA show that the symptoms 6 and 9 are sufficient for classification into the target number of mixtures. The plot of the minimum ICOMP score at each iteration for the KCS group 1 data set is shown in figure 6.2.

Now, we present the results for the group 2 medical data set. The information criteria scores generated are shown in table 6.13. Even in this case, the two mixture model has the minimum information criteria value for AIC, SBC, ICOMP and $ICOMP_{PEU}$. According to the maximum likelihood criteria, the six mixture model is considered to be good since it has the minimum lack of fit.

Table 6.12: KCS Group1: Best model & its ICOMP score for each iteration of the GA

Model	*# of Mixtures*	*ICOMP*
1 0 0 0 0 0 0 1 0 0 0	2	218.5677
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	2	141.1611
0 0 0 0 0 1 0 0 1 0	**2**	**141.1611**

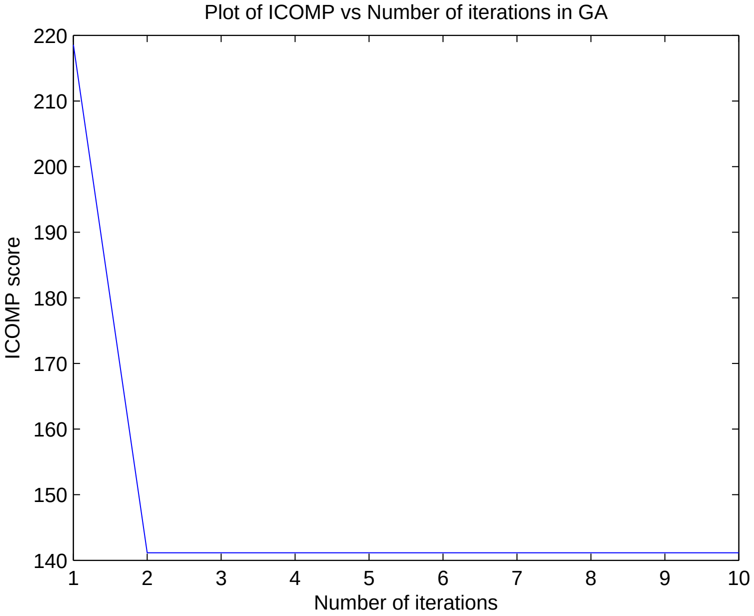


Figure 6.2: KCS Group1: Plot of ICOMP vs Number of iterations in GA

Table 6.13: KCS Group2: Lack of fit & information criteria values for mixtures 1 to 10

# of Mixtures	*Lackoffit*	*AIC*	*SBC*	*ICOMP*	*ICOMP (PEU)*
1	482.0649	512.0649	519.2006	512.4537	512.7868
2	397.9123	**460.9123**	**475.8973**	**474.2833**	**485.7394**
3	393.7224	489.7224	512.5567	503.8307	515.9185
4	391.7407	520.7407	551.4243	535.7644	548.6365
5	391.7313	553.7313	592.2642	568.932	581.9557
6	**391.2109**	586.2109	632.5931	601.3298	614.2835
7	391.2211	619.2211	673.4525	634.3989	647.4031
8	391.2379	652.2379	714.3187	667.4511	680.4856
9	391.2453	685.2453	755.1754	700.4152	713.4126
10	391.2447	718.2447	796.024	733.4081	746.3999

Table 6.14: KCS Group2: Mixing proportion estimates

Mixture	*Estimate*
$\hat{\pi_1}$	0.5870
$\hat{\pi_2}$	0.4130

Table 6.15: KCS Group2: Classification matrix

Mixture	*# of observations*
1	24
2	17

The estimated parameters for the two mixture model in the group 2 data set are given below in table 6.14 and table 6.16 respectively. Table 6.15 shows that out of 41 observations, the two mixture model classified 24 observations into mixture 1 and 17 observations into mixture 2. The confusion matrix for the above classification is given in table 6.17. We now present the results of our GA algorithm to determine which subset of the ten symptoms in the group 2 data set are sufficient for classification into the target number of mixtures.

The probability of mutation is 0.10. The population size is taken to be 50. The GA is run for 10 iterations. The following results are generated at the end of each iteration (table 6.18). The results of our GA show that the symptoms 7 and 10 are sufficient for classification into the target number of mixtures. The plot of the minimum ICOMP score at the end of each iteration for the KCS group 2 data set is shown in figure 6.3.

Table 6.16: KCS Group2: Probability estimates of the variables

Probability	$m = 1$	$m = 2$
p_1	0.8310	0.0000
p_2	0.8310	0.0000
p_3	0.7077	0.1753
p_4	0.6232	0.0000
p_5	0.4995	0.0578
p_6	0.4171	0.2929
p_7	0.3324	0.0000
p_8	0.3326	0.1769
p_9	0.2493	0.0000
p_{10}	0.2077	0.0000

Table 6.17: KCS Group2: Confusion matrix

Actual\Predicted	*Mixture1*	*Mixture2*
Mixture1	24	0
Mixture2	0	17

Table 6.18: KCS Group2: Best model & its ICOMP score for each iteration of the GA

Model	*# of Mixtures*	*ICOMP*
0 0 0 0 0 0 1 0 1 0	2	86.2947
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	2	76.2965
0 0 0 0 0 0 1 0 0 1	**2**	**76.2965**

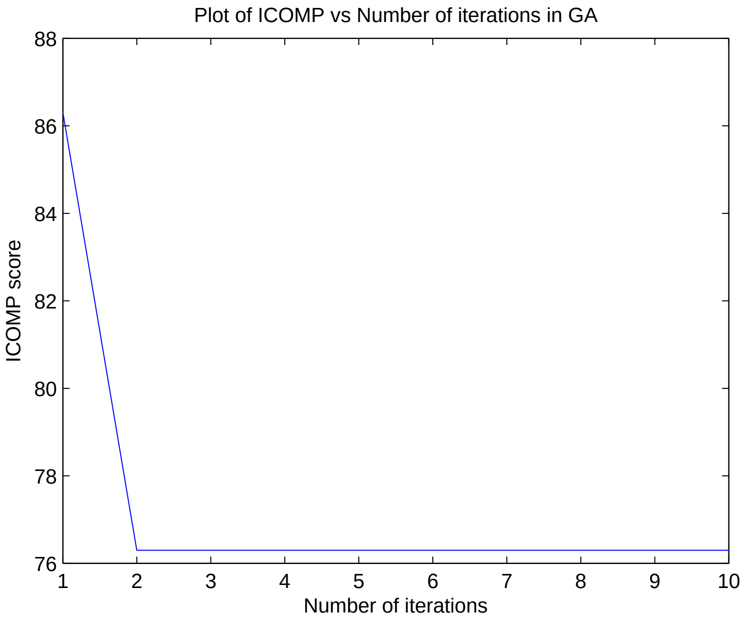


Figure 6.3: KCS Group2: Plot of ICOMP vs Number of iterations in the GA

6.2 Gifi - Multiple Regression

We use Gifi transformation on a mixed data set to transform the categorical predictor variables to a continuous space and then fit a multiple regression model to predict the continuous response. We show the results of our optimal scaling method (OSM) and linear combination method (LCM) on a simulated and two real data sets.

6.2.1 Simulation

We simulated a mixed data using the following protocol.

- $\varepsilon_1 \sim$ N(0,1).
- $\varepsilon_2 \sim$ N(0,1).
- $\varepsilon_3 \sim$ N(0,1).
- $x_1 = 10 + \varepsilon_1$.
- $x_2 = 10 + 0.3 \times \varepsilon_1 + 0.9539 \times \varepsilon_2$.
- $x_3 = 10 + 0.3 \times \varepsilon_1 + 0.5604 \times 0.9539 \times \varepsilon_2 + 0.8282 \times 0.9539 \times \varepsilon_3$.
- $x_4 \sim$ Bernoulli.
- $x_5 \sim$ Bernoulli.
- $x_6 \sim$ discrete U(1, 4).
- $x_7 \sim$ discrete U(1, 5).
- $x_8 \sim$ discrete U(1, 4).
- $y = -8 + x_1 + 0.3 \times x_3 + 0.5 \times x_5 + 0.4 \times x_6 + 0.6 \times x_8$.

We ran multiple regression procedure on this mixed data with y as response and $x_1 - x_8$ as predictors. We included the intercept in this model. We use GA for variable selection with the following parameters: maximum iterations - 100, population size - 20, probability of crossover - 0.75, probability of mutation - 0.10, crossover type - uniform, fitness function - $ICOMP_{C1}$. The variables selected are $Intercept, x_1, x_3, x_4, x_5, x_6, x_7, x_8$. These selected variables seems that the model in the original mixed space is over-fitted. Now, we transform this mixed space to a pure continuous space using our Gifi system and run multiple regression procedure on the transformed continuous space. The variables selected using OSM procedure are $Intercept, x_1, x_3, x_5, x_6, x_8$. The variables selected using the OSM procedure are the right set of predictors. The variables selected using LCM procedure are $Intercept, x_1, x_3, x_5, x_6$. The variables selected using the LCM procedure are almost the same as the ones selected by OSM procedure except for x_8. This can be accounted for the loss of information using the LCM procedure.

6.2.2 Beta-Carotene Data

Beta-Carotene Data: Linear Combination Method

Observational studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with increased risk of developing certain types of cancer. However, relatively few studies have investigated the determinants of plasma concentrations of these micronutrients. A cross-sectional study was designed to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of retinol, beta-carotene and other carotenoids. Study subjects (N = 315) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous. We display the data for only two of the analytes (BETAPLASMA and RETPLASMA).

Variable names:

- AGE: Age (years)
- SEX: Sex (1=Male, 2=Female).
- SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker)
- QUETELET: Quetelet ($weight/(height^2)$)
- VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)

- CALORIES: Number of calories consumed per day.
- FAT: Grams of fat consumed per day.
- FIBER: Grams of fiber consumed per day.
- ALCOHOL: Number of alcoholic drinks consumed per week.
- CHOLESTEROL: Cholesterol consumed (mg per day).
- BETADIET: Dietary beta-carotene consumed (mcg per day).
- RETDIET: Dietary retinol consumed (mcg per day)
- BETAPLASMA: Plasma beta-carotene (ng/ml)
- RETPLASMA: Plasma Retinol (ng/ml)

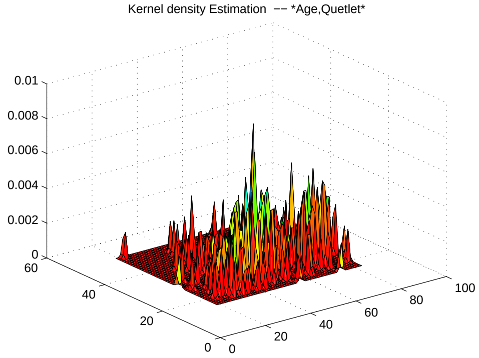
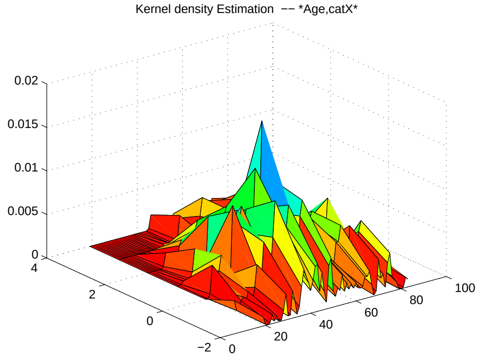
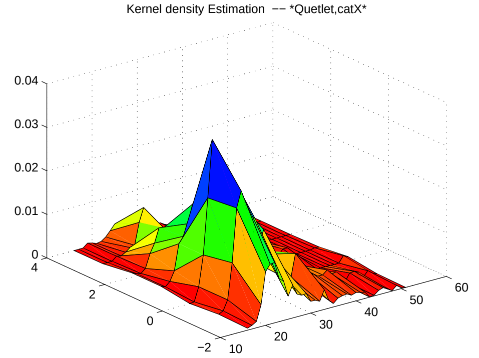
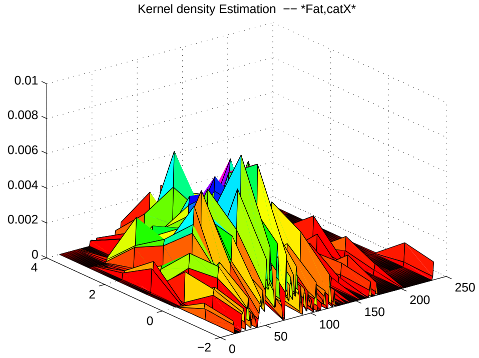
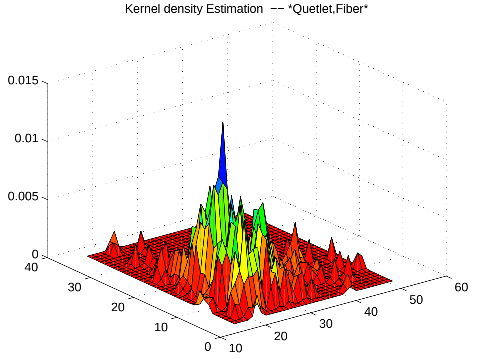
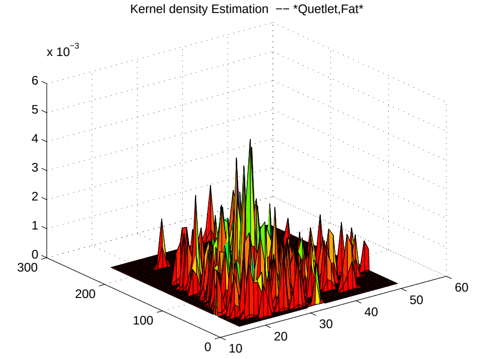
This data has not been published yet but a related reference is [Nierenberg et al., 1989].

Since the variables SEX, SMOKSTAT, and VITUSE are categorical, we use the Gifi transformation to generate an optimal weight(score) vector that is used for transforming the categorical space to the continuous space. Some of the pair-wise kernel density estimates of this data in the Gifi space is shown in table 6.2.2. In this work, we used gaussian kernel with bandwidth, $h = 0.5$.

We fit a multiple regression model with RETPLASMA as the dependent variable and the variables AGE, SEX, SMOKSTAT, QUETELET, VITUSE, CALORIES, FAT, FIBER, ALCOHOL, CHOLESTROL, BETADIET and RETDIET as independent variables. We also include the intercept term in this model. We assume that the residuals are normally distributed. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The best set of variables selected by GA and its associated information score ($ICOMP_{C1}$)is given by

Table 6.19: Kernel Density Estimate of Beta-Carotene Data in the Gifi space

KDE1	KDE2
 Kernel density Estimation -- *Age,Quetlet* 	 Kernel density Estimation -- *Age,catX*
 Kernel density Estimation -- *Quetlet,catX* 	 Kernel density Estimation -- *Fat,catX*
 Kernel density Estimation -- *Quetlet,Fiber* 	 Kernel density Estimation -- *Quetlet,Fat*

Model: Intercept, Age, Sex, Smokstat, Quetlet, Fat, Alcohol, Retdiet

Information criteria score: 4233.376. The parameter estimates for the above set of variables is given by

$$\beta = \begin{bmatrix} 500.7981 \\ 2.5039 \\ 1.0760 \\ -0.5323 \\ -0.2685 \\ -0.0118 \\ 37.2665 \end{bmatrix}$$

The RMSE for this model is 200.3348. The optimal weights(scores) associated with the categories of the variables Sex and Smokestat are

$$wSex = \begin{bmatrix} 1.7298 \\ -0.2661 \end{bmatrix} \qquad wSmokeStat = \begin{bmatrix} -0.6169 \\ 0.5241 \\ 0.8508 \end{bmatrix}$$

Hence, the regression equation can be written as

$$\begin{aligned} RETPLASMA &= 500.7981 + 2.5039 \times Age + 1.0760 \times Quetlet - \\ &= 0.5323 \times Fat - 0.2685 \times Alcohol - 0.0118 \times Retdiet + \\ &= 37.2665 \times catX \end{aligned}$$

where catX is the linear combination of the weights(scores) of the categories of the variables SEX and SMOKESTAT respectively.

The model given by the stepwise variable selection using NCSS software is given by

Model: Intercept, Age, Sex

and the RMSE computed by NCSS for the above model is 202.99. The RMSE computed by NCSS on the data in the original mixed space for the model selected by GA and ICOMP in the Gifi space is 201.041. In this example, the way to analyze the efficiency of this

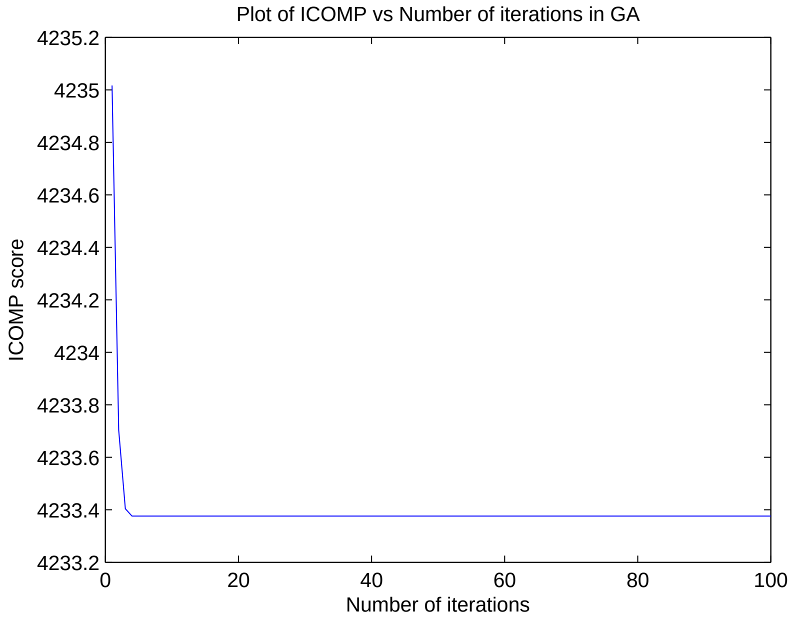


Figure 6.4: Beta-Carotene (RetPlasma): Plot of ICOMP vs Number of iterations in GA

transformation is by general intuition. The variable that is very much related to tumor development is SmokeStat. This variable is not picked up by the model in the original mixed space whereas it is picked up by the model in the Gifi space.

For instance, if the categorical variable *Smokestat* takes value 1 and the categorical variable *Sex* takes value 0. The corresponding weight associated with a value of 1 for the categorical variable *Smokestat* is 0.5241 and the corresponding weight associated with a value of 0 for the categorical variable *Sex* is 1.7298. Therefore, the value of *catX* would be

$$0.5241 + 1.7298 = 2.2539$$

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.4. The plot matrix of the best set of predictors selected by GA and ICOMP is given in figure 6.7. The plot matrix is in the order of the predictors reported in the model i.e., Age, Quetlet, Fat, Alcohol, Retdiet and catX where catX is the linear combination of the categorical variables Sex and Smokestat in the Gifi space. The set of variables selected by AIC with same set of GA parameters is given by

Model: Intercept, Age, Sex, Smokstat, Fat

The AIC score for this model is 4241.647. The estimated parameters for the model given by AIC are given by

$$\beta = \begin{bmatrix} 527.5606 \\ 2.4593 \\ -0.6243 \\ 36.5431 \end{bmatrix}$$

Since ICOMP is more consistent in choosing the right model, we consider the model selected by ICOMP as our best fitting model for this data.

Beta-Carotene Data: Optimal Scaling Method

The data is transformed to the Gifi space and the categorical variables are optimally scaled. We fit a multiple regression model with RETPLASMA as the dependent variable and the variables in the Gifi space AGE, SEX, SMOKSTAT, QUETELET, VITUSE, CALORIES, FAT, FIBER, ALCOHOL, CHOLESTROL, BETADIET and RETDIET as independent variables. We also include the intercept term in this model. We assume that the residuals are normally distributed. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The best set of variables selected by GA and its associated information score is given by

Model: Intercept, Age, Sex, Smokstat, Vituse

The ICOMP score for this model is 4244.857. The parameter estimates for the above set of variables is given by

$$\beta = \begin{bmatrix} 469.2317 \\ 2.6634 \\ 41.0245 \\ 26.9100 \\ -22.4476 \end{bmatrix}$$

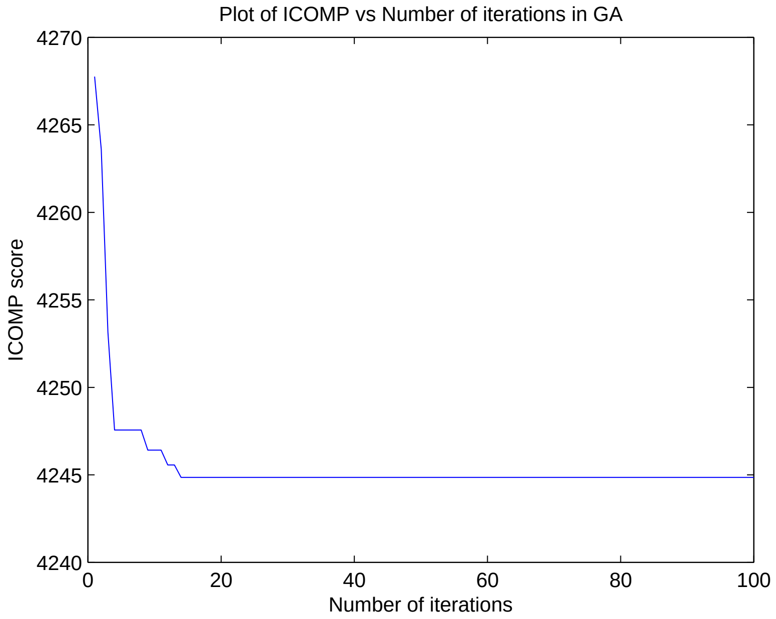


Figure 6.5: Beta-Carotene (RetPlasmaOS): Plot of ICOMP vs Number of iterations in GA

Hence the regression equation in the Gifi space can be given by

$$\begin{aligned} RETPLASMA &= 469.2317 + 2.6634 \times Age + 41.0245 \times Sex \\ &\quad +26.9100 \times Smokstat - 22.4476 \times Vituse \end{aligned}$$

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.5. The set of variables selected by AIC with same set of GA parameters is given by

Intercept, Age, Sex, Smokstat, Fat

The AIC score for this model is 4243.2061. The estimated parameters for the above model is given by

$$\beta = \begin{bmatrix} 537.5733 \\ 2.2938 \\ 45.5978 \\ 27.1234 \\ -0.6466 \end{bmatrix}$$

Since ICOMP is more consistent in choosing the right model, we choose the model selected by ICOMP as our best fitting model.

6.2.3 Cars Data

Cars Data: Linear Combination Method

This data is taken from JSE (Journal of Statistical Education) data archive. It contains new car specifications for the year 2004. It contains 387 observations on 19 variables. There are no missing values in this data. The description of the 19 variables is briefly listed below:

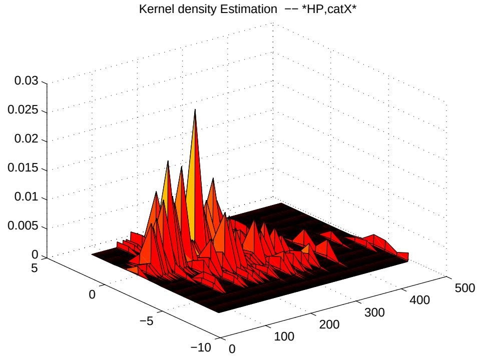
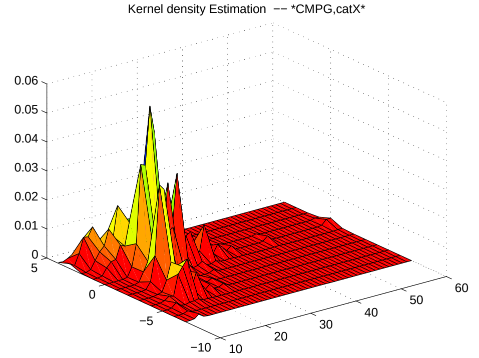
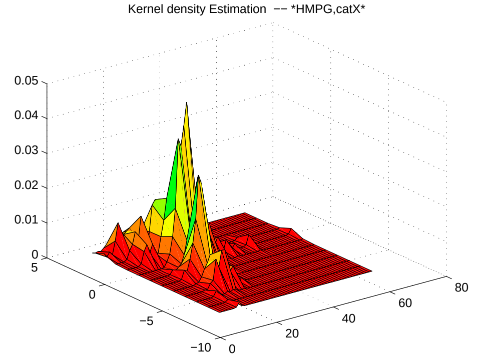
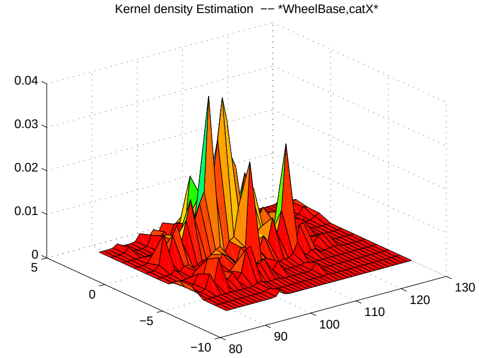
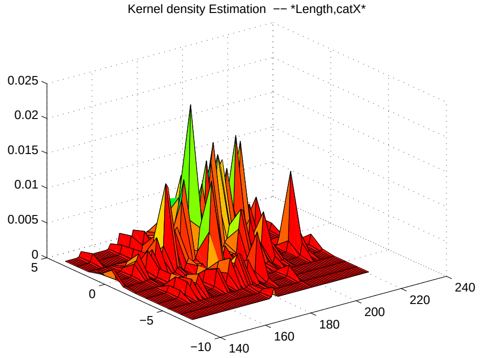
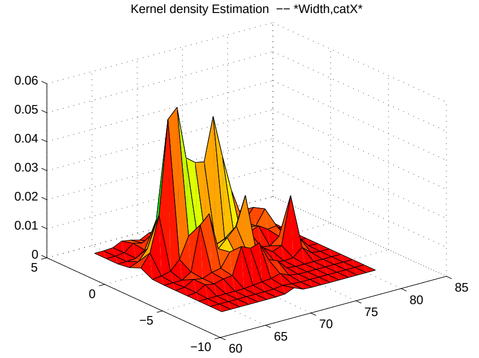
- Sports Car (1=yes, 0=no)
- SUV: Sport Utility Vehicle (1=yes, 0=no)
- Wagon (1=yes, 0=no)
- Minivan (1=yes, 0=no)
- Pickup (1=yes, 0=no)
- AWD: All-Wheel Drive (1=yes, 0=no)
- RWD: Rear-Wheel Drive (1=yes, 0=no)
- SRP: Suggested Retail Price, what the manufacturer thinks the vehicle is worth, including adequate profit for the automaker and the dealer (U.S. Dollars)
- DC: Dealer Cost (or ”invoice price”), what the dealership pays the manufacturer (U.S. Dollars)

- Engine size (liters)
- NumCylinders: Number of Cylinders
- HP: Horsepower
- CMPG: City Miles Per Gallon
- HMPG: Highway Miles Per Gallon
- Weight (Pounds)
- Wheel Base (inches)
- Length (inches)
- Width (inches)

The variables Sports Car, SUV, Wagon, Minivan, Pickup, AWD, RWD, NumCylinders are categorical and the variables SRP/DC, Engine size, HP, CMPG, HMPG, Weight, Wheel Base, Length and Width are continuous variables. The variables SRP/DC, CMPG and HMPG can be considered to be the most obvious choice for the response variable(s). We use the Gifi transformation on the categorical predictor variables to generate an optimal weight(score) vector that is used for transforming the categorical space to the continuous space. Some of the pair-wise kernel density estimates of this data in the Gifi space is shown in table 6.2.3.

We fit a multiple regression model to the cars data with SRP as the response variable and the variables Sports Car, SUV, Wagon, Minivan, Pickup, AWD, RWD, NumCylinders, Engine size, HP, CMPG, HMPG, Weight, Wheel Base, Length and Width as predictor variables. We also include the intercept term in this model. We assume that the residuals are normally distributed. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

Table 6.20: Kernel Density Estimate of Cars Data in the Gifi space

KDE1	KDE2
	
	
	

The best set of variables selected by GA and its associated information score is given by

Model: Intercept, SUV, Wagon, RWD, Engine size, NumCylinders, HP, HMPG, Weight, Wheelbase, Width

Information criteria score: 8199.6978. The parameter estimates for the above set of variables is given by

$$\beta = \begin{bmatrix} 47055 \\ -1942 \\ 222 \\ 681 \\ 15 \\ -765 \\ -621 \\ -3523 \end{bmatrix}$$

with RMSE of 9592.1. Hence the regression equation can be given by

$$\begin{aligned} SRP &= 47055 - 1942 \times Enginesize + 222 \times HP + \\ &= 681 \times HMPG + 15 \times Weight - 765 \times WheelBase - \\ &= 621 \times Width - 3523 \times catX \end{aligned}$$

where catX is the linear combination of the weights of the categories of the variables SUV, Wagon, RWD and NumCylinders. The weight vector associated with the categories of the variable SUV, Wagon, Minivan, RWD, and NumCylinders are given by

$$wSUV = \begin{bmatrix} -0.2595 \\ 1.4426 \end{bmatrix} \quad wWagon = \begin{bmatrix} -0.0174 \\ 0.2152 \end{bmatrix} \quad wMinivan = \begin{bmatrix} -0.0148 \\ 0.2721 \end{bmatrix}$$

$$wRWD = \begin{bmatrix} 0.4403 \\ -1.37256 \end{bmatrix} \qquad wNumCylinders = \begin{bmatrix} 0.0433 \\ 0.1204 \\ 1.1637 \\ 0.0239 \\ -0.2795 \\ -3.1550 \end{bmatrix}$$

The model given by the stepwise (backward) regression method using JMP software is given by

Model: Intercept, Minivan, RWD, Engine size, NumCylinders, HP, HMPG, Weight,
Wheel base, Length, Width

and the RMSE computed by JMP for the above model is 9741.286. The RMSE computed by JMP for the model selected by Gifi is 9778.26. This shows that the model fitting might be better in the Gifi space than the original space when there are many categorical variables. In this case, the categorical variables that are directly related to the SRP are SUV, AWD/RWD, and NumCylinders. Clearly, SUV is not picked by the model in the original mixed space whereas it is picked up by the model in the Gifi space.

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.6. A plot matrix of the best predictors in the Gifi space is shown in figure 6.8. The plot matrix is in the order of the predictors EngSize, HP, HMPG, Weight, WheelBase, Width and catX where catX is the linear combination of the categories of the categorical variables SUV, Wagon, RWD and NumCylinders in the Gifi space.

The model selected by AIC with the same set of GA parameters is given by

Model: Intercept, SUV, Wagon, RWD, Engine size, NumCylinders, HP, HMPG, Weight,
Wheel base, Width

The AIC score for this model is 8210.8271. Since ICOMP is more consistent in choosing the right model, we consider the model selected by ICOMP as the best fitting model for

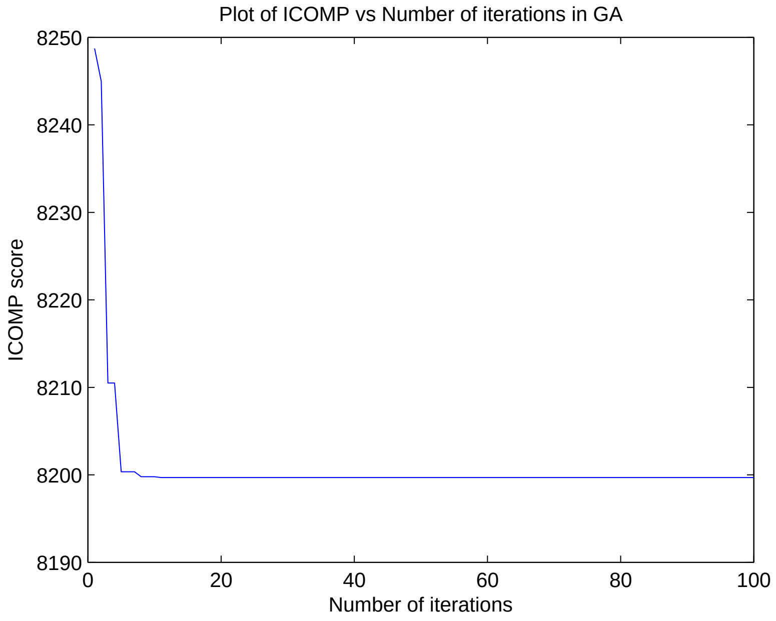


Figure 6.6: Cars Data (SRP): Plot of ICOMP vs Number of iterations in GA

this data set.

Cars Data: Optimal Scaling Method

The data is transformed to the Gifi space and the categorical variables are optimally scaled. We fit a multiple regression model to the cars data with SRP as the response variable and the variables Sports Car, SUV, Wagon, Minivan, AWD, RWD, NumCylinders, Engine size, HP, CMPG, HMPG, Weight, Wheel Base, Length and Width as predictor variables. We also include the intercept term in this model. We assume that the residuals are normally distributed. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The best set of variables selected by GA and its associated information score is given by

SportsCar, SUV, AWD, RWD, EngSize, NumCylinders, HP, HMPG, Weight, WheelBase

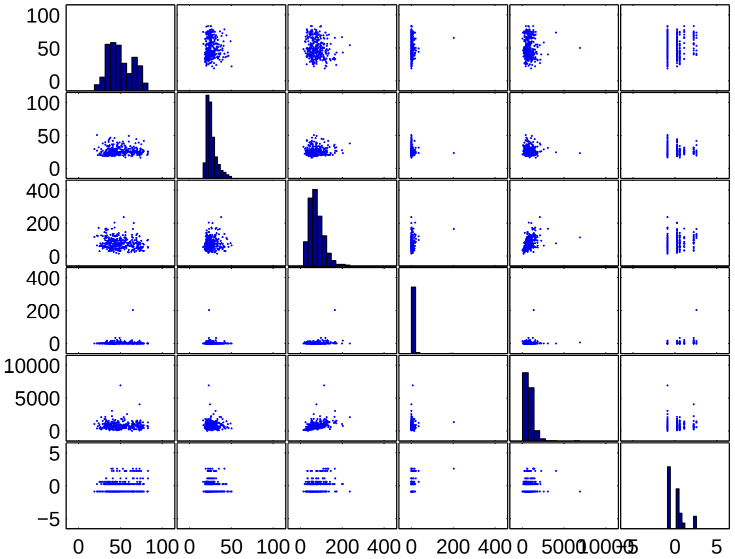


Figure 6.7: Beta-Carotene: Plot matrix of the best predictors in the model RetPlasma

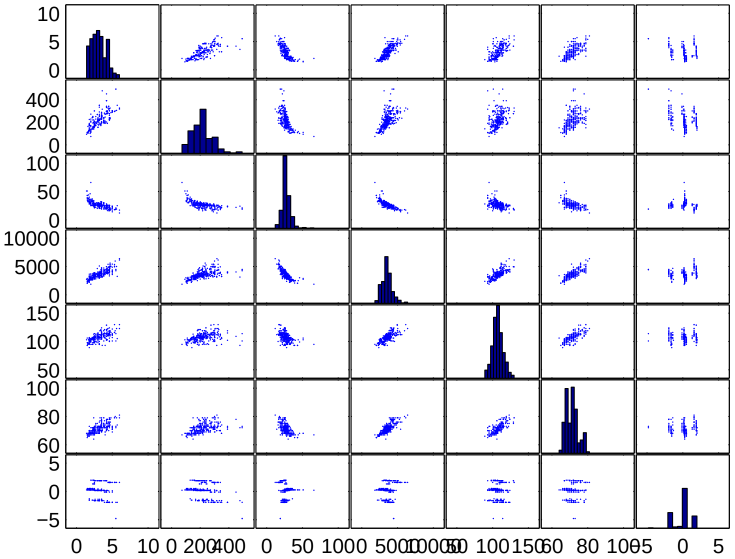


Figure 6.8: Cars(SRP) Data: Plot Matrix of the predictors in the Gifi space

The ICOMP score for this model is 8232.5139. The parameter estimates for the above model is given by

$$\beta = \begin{bmatrix} -1045.9 \\ -2271.5 \\ 1317.9 \\ -3745.8 \\ -3194.8 \\ -7070.1 \\ 225.0 \\ 756.5 \\ 11.7 \\ -626.1 \end{bmatrix}$$

Hence the regression equation in the Gifi space is given by

$$\begin{aligned} SRP &= -1045.9 \times SportsCar - 2271.5 \times SUV + \\ &\quad 1317.9 \times AWD - 3745.8 \times RWD - 3194.8 \times EngSize \\ &\quad -7070.1 \times NumCylinders + 225 \times HP + \\ &\quad 756.5 \times HMPG + 11.7 \times Weight - 626.1 \times WheelBase \end{aligned}$$

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.9. The set of predictors selected by AIC is given by

Intercept, SUV, RWD, EngSize, NumCylinders, HP, HMPG, Weight, WheelBase, Width

The AIC score is: 8212.7737. Since ICOMP is more consistent in choosing the right model, we consider the model selected by ICOMP as our best fitting model.

6.3 Gifi - Binary Logistic Regression

We use Gifi transformation on a mixed data set to transform the categorical predictor variables to a continuous space and then fit a binary logistic regression model to predict

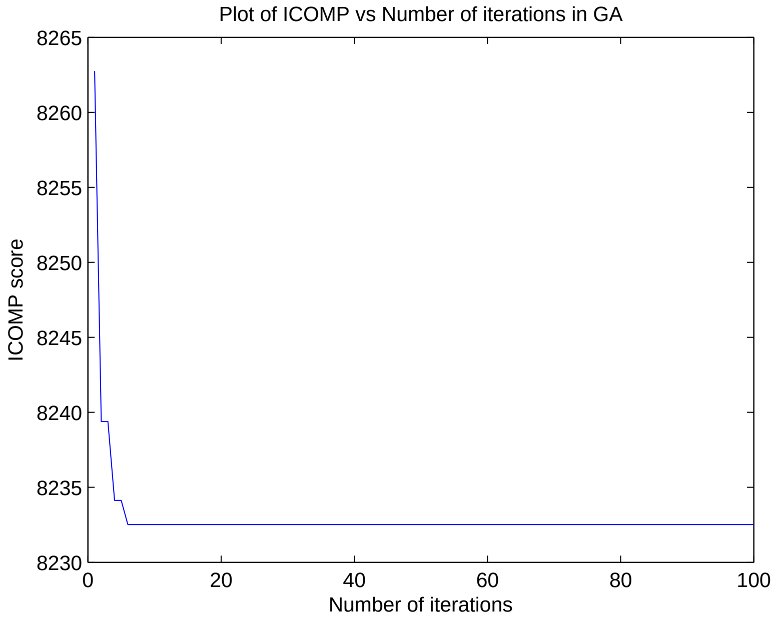


Figure 6.9: Cars Data (SRPOS): Plot of ICOMP vs Number of iterations in GA

the binary response. We show the results of our algorithm on two real data sets.

6.3.1 ICU Data

ICU Data: Linear Combination Method

The data consist of 200 subjects from a larger study on the survival of patients following admission to an adult intensive care unit (ICU). The study used logistic regression to predict the probability of survival for these patients until their discharge from the hospital. The dependent variable is the binary variable Vital Status (STA). Nineteen possible predictor variables, both discrete and continuous, were also observed.

Variable names:

- ID: ID number of the patient
- STA: Vital status (0 = Lived, 1 = Died)
- AGE: Patient's age in years
- SEX: Patient's sex (0 = Male, 1 = Female)

- RACE: Patient's race (1 = White, 2 = Black, 3 = Other)
- SER: Service at ICU admission (0 = Medical, 1 = Surgical)
- CAN: Is cancer part of the present problem? (0 = No, 1 = Yes)
- CRN: History of chronic renal failure (0 = No, 1 = Yes)
- INF: Infection probable at ICU admission (0 = No, 1 = Yes)
- CPR: CPR prior to ICU admission (0 = No, 1 = Yes)
- SYS: Systolic blood pressure at ICU admission (in mm Hg)
- HRA: Heart rate at ICU admission (beats/min)
- PRE: Previous admission to an ICU within 6 months (0 = No, 1 = Yes)
- TYP: Type of admission (0 = Elective, 1 = Emergency)
- FRA: Long bone, multiple, neck, single area, or hip fracture (0 = No, 1 = Yes)
- PO2: PO2 from initial blood gases ($0 = > 60$, $1 = 60$)
- PH: PH from initial blood gases ($0 = 7.25$, $1 < 7.25$)
- PCO: PCO2 from initial blood gases ($0 = 45$, $1 = > 45$)
- BIC: Bicarbonate from initial blood gases ($0 = 18$, $1 = < 18$)
- CRE: Creatinine from initial blood gases ($0 = 2.0$, $1 = > 2.0$)
- LOC: Level of consciousness at admission (0 = no coma or stupor, 1 = deep stupor, 2 = coma)

This data set is first run using the logistic regression in NCSS. It resulted in giving a warning message stating that the maximum likelihood criteria failed to converge. The warning message given by the NCSS software tells us that the usual statistical software fails to come up with an accurate or a near accurate model when the data is of the mixed type. The results that are given by NCSS shows that only the parameters RACE, CAN and LOC are significant. These results are not be believed since the software failed to get the maximum likelihood routine to converge. Even the software JMP produced a similar result. Hence, we can say that the present commercial software's fail to produce good results on such data sets.

Now, we show the results of our Gifi system on this data set. The variables AGE, SYS and HRA are continuous predictors and the variables SEX, RACE, SER, CAN, CRN, INF, CPR, PRE, TYP, FRA, PO2, PH, PCO, BIC, CRE and LOC are categorical predictors. STA is a binary response variable. We first apply the Gifi transformation on the categorical predictors to generate an optimal weight(score) vector that is used for transforming the categorical space to a pure continuous space. A plot matrix of the data in the Gifi space is shown in figure 6.12. The plot matrix is in the order of the predictors, AGE, SYS, HRA, catX where catX is the linear combination of the categories of the categorical variables in the Gifi space. Some of the pair-wise kernel density estimates of this data in the Gifi space is given in table 6.3.1.

We fit a binary logistic regression in the Gifi space.The input model includes an intercept term. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{IFIM}$ is used as the fitness function.

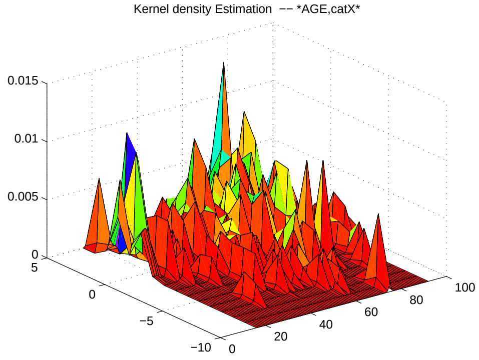
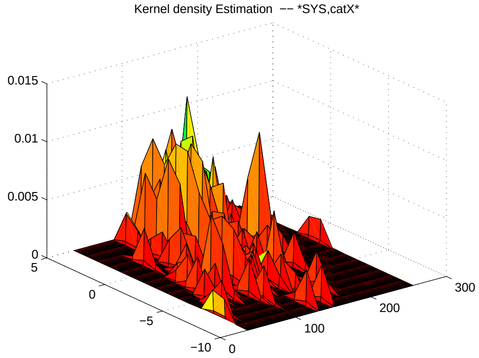
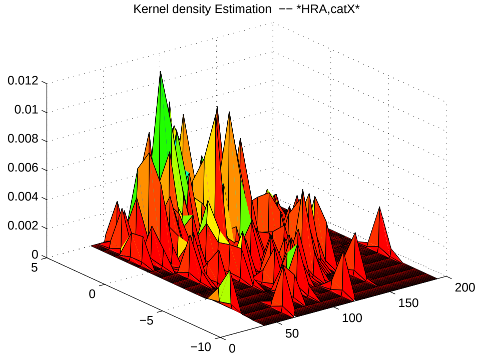
The following model is selected by GA with above input parameters.

$$\text{Model} = \text{Intercept, RAC, INF, TYP, FRA, LOC}$$

Information Criterion Score: 164.102. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.10. The parameters INF, TYP, FRA and LOC are all related to the seriousness of the patients' condition.

The confusion matrix is given in table 6.22. The prediction accuracy is 84% and the

Table 6.21: Kernel Density Estimate of ICU Data in the Gifi space

KDE1	KDE2
	
	

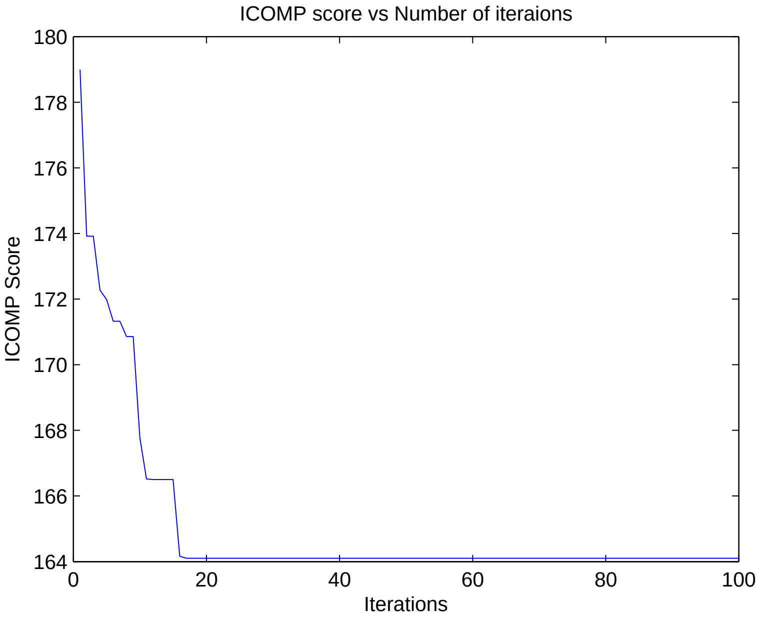


Figure 6.10: ICU Data (BLR): Plot of ICOMP vs Number of iterations in GA

Table 6.22: Classification matrix

	Lived(0)	*Died(1)*
Lived(0)	158	30
Died(1)	2	10

error rate is 16%. The estimated parameters are given by

$$\beta = \begin{bmatrix} -1.7573 \\ 1.3035 \end{bmatrix}$$

where -1.7573 is the coefficient corresponding to the intercept term and 1.3035 is the coefficient corresponding to the categorical terms in the model. The weights associated with the variable RAC, INF, TYP, FRA, and LOC are given by

$$wRAC = \begin{bmatrix} 0.0178 \\ -0.1572 \\ -0.0757 \end{bmatrix} \qquad wINF = \begin{bmatrix} -0.3867 \\ 0.5341 \end{bmatrix} \qquad wTYP = \begin{bmatrix} -0.9797 \\ 0.3532 \end{bmatrix}$$

$$wFRA = \begin{bmatrix} 0.0140 \\ -0.1725 \end{bmatrix} \qquad wLOC = \begin{bmatrix} -0.1087 \\ 0.4921 \\ 1.7652 \end{bmatrix}$$

For instance, a value of 1 for RAC, 1 for INF, 1 for TYP, 1 for FRA, and 1 for LOC would yield

$$0.0178 - 0.3867 - 0.9797 + 0.0140 - 0.1087 = -1.4433$$

Therefore, the predicted probability, p, is given by

$$\begin{aligned} p &= \frac{e^{-1.7573-1.3035\times1.4433}}{1+e^{-1.7573-1.3035\times1.4433}} \\ &= 0.0256 \end{aligned}$$

Since the value of p is less than 0.5, the category of STA predicted is 0.

The data in the original space is run for binary logistic regression using GA and $ICOMP_{IFIM}$ as the model selection criteria. The best model selected by GA is given by

Model: AGE, RAC, SER, CAN, CPR, SYS, PRE, TYP, PCO, LOC

The model fitted in the Gifi space is much better in terms of sparsity. The model selected by AIC is given by

Model: Intercept, AGE, RAC, TYP, LOC

The AIC score for this model is 160.23. Since ICOMP is more consistent in choosing the right model, we consider the model selected by ICOMP as the best fitting model.

ICU Data: Optimal Scaling Method

The data is transformed to the Gifi space and the categorical variables are optimally scaled. We fit a binary logistic regression model with STA as the binary response variable and AGE, SEX, RACE, SER, CAN, CRN, INF, CPR, SYS, HRA, PRE, TYP, FRA, PO2, PH, PCO, BIC, CRE and LOC as predictor variables. The input model includes an intercept term. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{IFIM}$ is used as the fitness function.

The following model is selected as best by the GA with the above input parameters.

Intercept, AGE, TYP, LOC

The information criteria score: 166.6826. The estimated parameters for this model is given by

$$\beta = \begin{bmatrix} -3.6453 \\ 0.0325 \\ -1.6342 \\ -1.8151 \end{bmatrix}$$

The confusion matrix is given in table 6.23. The prediction accuracy is 83.5% and the error rate is 16.5%. The best value of the above fitness function at the end of each iteration of the GA process is shown in figure 6.11.

Table 6.23: Classification matrix

	Lived(0)	*Died(1)*
Lived(0)	158	31
Died(1)	2	9

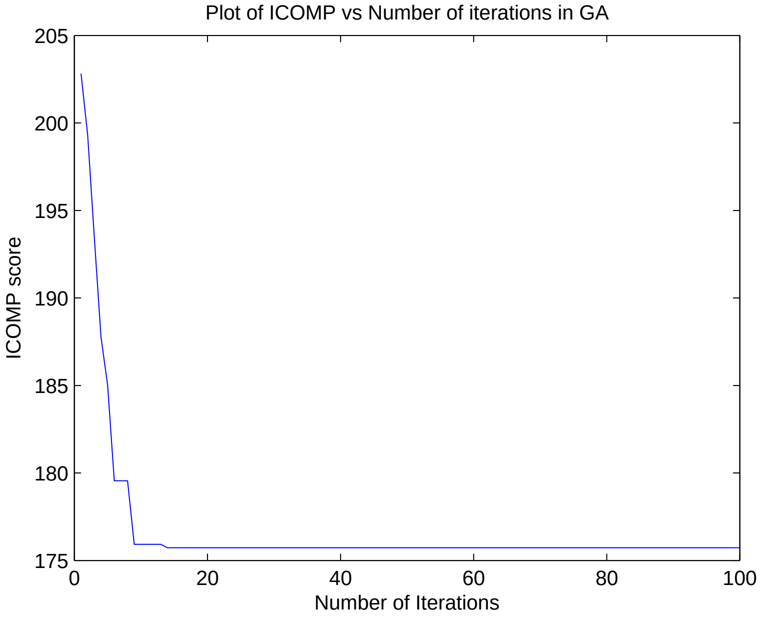


Figure 6.11: ICU Data (BLROS): Plot of ICOMP vs Number of iterations in GA

6.3.2 Prostate Cancer Data

Prostate Cancer Data: Linear Combination Method

This data is taken from [Hosmer and Lemeshow, 2000]. The data contains 380 observations on 9 variables with 4 missing observations. The missing observations were deleted leaving 376 observations for the analysis. The variables are briefly listed below:

- CAPSULE - Tumor Penetration of Prostatic Capsule, (0 - No Penetration, 1 - Penetration).
- AGE
- RACE (1 - white, 2 - black)
- DPROS - Results of the Digital Rectal Exam (1 - No Nodule, 2 - Unilobar Nodule (Left), 3 - Unilobar Nodule (Right), 4 - Bipolar Nodule
- DCAPS - Detection of Capsular Involvement in Rectal Exam (1 - No, 2 - Yes)
- PSA - Prostatic Specific Antigen Value (mg/ml)
- VOL - Tumor Volume Obtained from Ultrasound (cm3)
- GLEASON - Total Gleason Score (0 - 10)

The data on these 376 observations has been transformed to the Gifi space. A plot matrix of the data in the Gifi space is shown in figure 6.13. The plot matrix is in the order of the variables AGE, PSA, VOL, catX where catX is the linear combination of the categories of the categorical variables in the Gifi space. Some of the pairwise kernel density estimates of this data is given in table 6.3.2.

A binary logistic regression model is fit to the data in the Gifi space with CAPSULE as the response variable and RACE, DPROS, DCAPS, GLEASON, AGE, PSA and VOL as predictor variables. The input model includes an intercept term. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{IFIM}$ is used

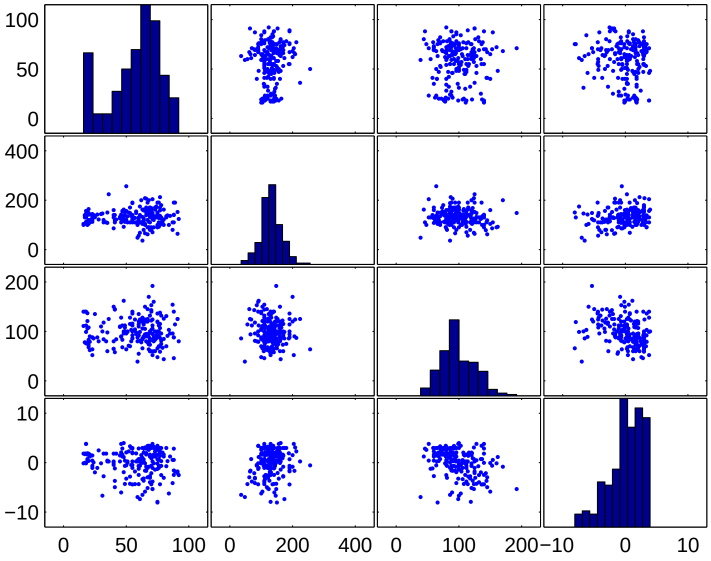


Figure 6.12: ICU Data: Plot Matrix of the data in the Gifi space

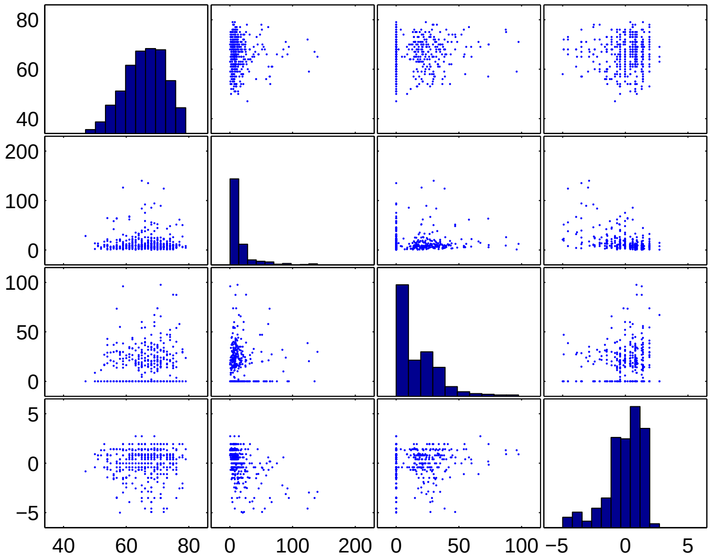
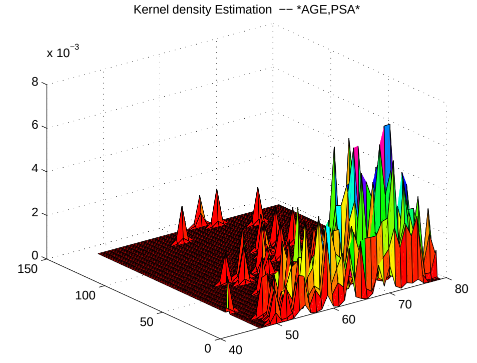
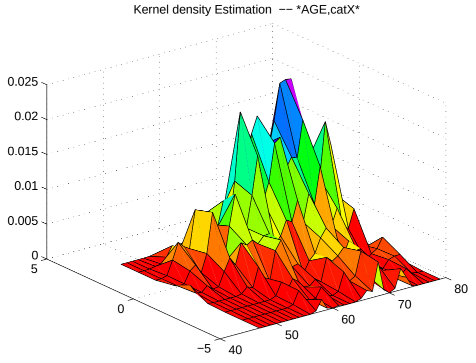
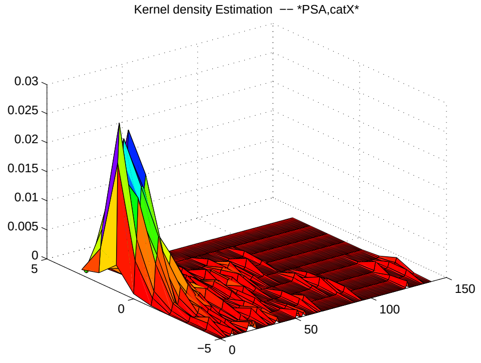


Figure 6.13: PCD Data: Plot Matrix of the data in the Gifi space

Table 6.24: Kernel Density Estimate of Prostate Cancer Data in the Gifi space

KDE1	KDE2
	
	

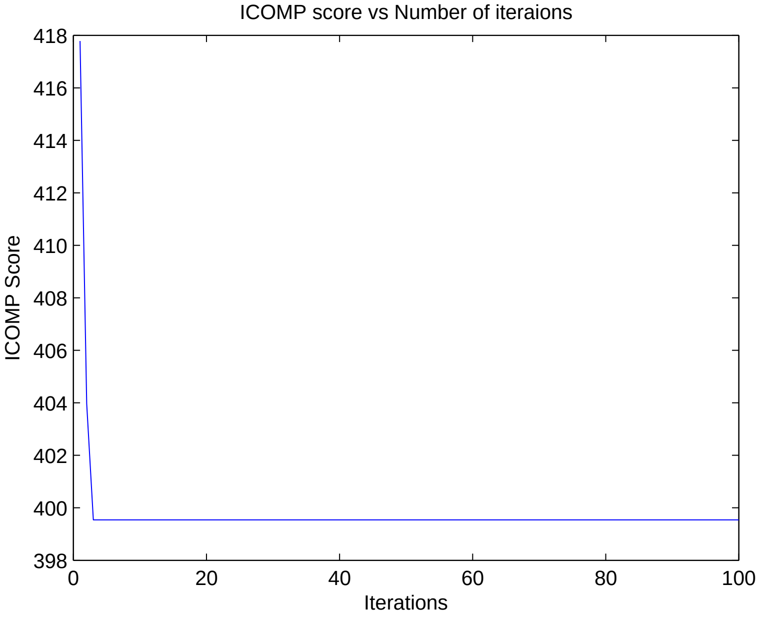


Figure 6.14: PCD (BLR): Plot of ICOMP vs Number of iterations in GA

Table 6.25: Classification matrix

CAPSULE	*No Penetration*	*Penetration*
No Penetration	185	62
Penetration	40	89

as the fitness function.

The following model is selected as best model by the GA process with above input parameters.

$$\text{Model} = \text{Intercept, DPROS, PSA, GLEASON}$$

Information Criterion Score: 399.5404. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.14. The confusion matrix is given in 6.25. The prediction accuracy is 72.87% and the error rate is 27.12%. The data in the original mixed space is run for binary logistic regression using NCSS. The only variables found significant are DPROS and PSA resulting in a prediction accuracy of 69.737%. The confusion matrix is given in table 6.26. For this data set, the model fitting in Gifi space is better than the model fitting in the original mixed space.

Table 6.26: Classification matrix

CAPSULE	*No Penetration*	*Penetration*
No Penetration	197	30
Penetration	85	68

Table 6.27: Classification matrix

CAPSULE	*No Penetration*	*Penetration*
No Penetration	193	56
Penetration	32	95

Prostate Cancer Data: Optimal Scaling Method

The data is transformed to the Gifi space and the categorical variables are optimally scaled. We fit a binary logistic regression model with CAPSULE as the response variable and RACE, DPROS, DCAPS, GLEASON, AGE, PSA and VOL as predictor variables. The input model includes an intercept term. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{IFIM}$ is used as the fitness function.

The following model is selected as best model by the GA process with above input parameters.

AGE, DPROS, PSA, GLEASON

The information criteria score: 403.5058. The parameter estimates for this model is given by

$$\beta = \begin{bmatrix} -0.0121 \\ -0.7721 \\ 0.0237 \\ -1.2747 \end{bmatrix}$$

The confusion matrix is given in table 6.27. The prediction accuracy is 76.595% and the error rate is 23.4%. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.15.

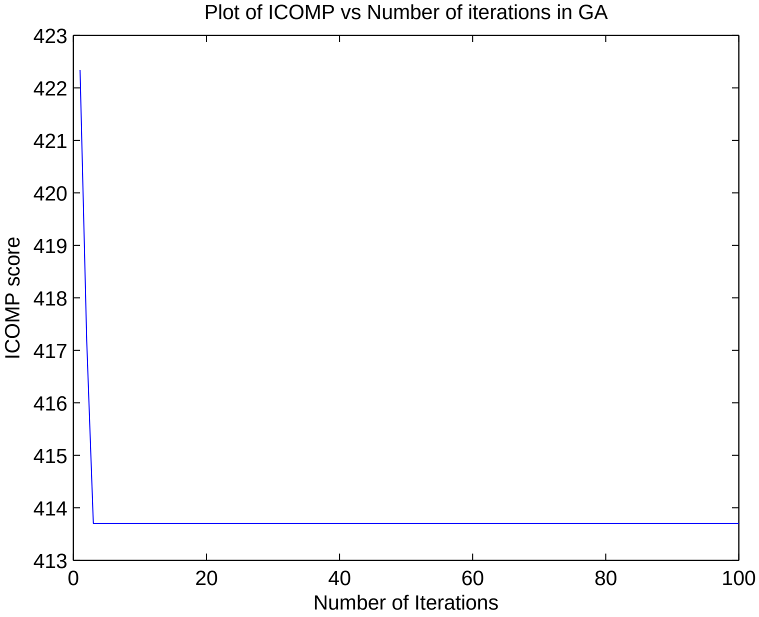


Figure 6.15: PCD (BLROS): Plot of ICOMP vs Number of iterations in GA

6.4 Gifi - Multivariate Regression

We use Gifi transformation on a mixed data set to transform the categorical predictor variables to a pure continuous space and then fit a multivariate regression model to predict a set of continuous responses. We show the results of our algorithm on two real data sets.

6.4.1 Beta-Carotene

Beta-Carotene – Linear Combination Method

We fit a multivariate regression model for the Beta-Carotene data with BETAPLASMA and RETPLASMA as dependent variables and AGE, SEX, SMOKSTAT, QUETELET, VITUSE, CALORIES, FAT, FIBER, ALCOHOL, CHOLESTROL, BETADIET and RETDIET as independent variables. Since the variables SEX, SMOKSTAT, and VITUSE are categorical, we use the Gifi transformation to come up with an optimal weight vector that is used for transforming the categorical space to a pure continuous space. We assume that the residuals are multivariate normally distributed. We also include the intercept term in this model. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The best set of variables selected by GA and its associated information score is given by

Model: Intercept, Sex, Smokstat, Vituse

Information criteria score: 14785.0949.

The parameter estimates for the above set of variables is given by

$$\beta = \begin{bmatrix} 189.8921 & 602.7905 \\ -29.7136 & 18.4572 \end{bmatrix}$$

A new observation vector Y_i can be predicted from $X_i \times Beta$.

The optimal weights associated with the categories of the variable Sex, Smokestat and Vituse are given by

$$wSex = \begin{bmatrix} 1.7298 \\ -0.2661 \end{bmatrix} \quad wSmokeStat = \begin{bmatrix} -0.6169 \\ 0.5241 \\ 0.8508 \end{bmatrix} \quad wVituse = \begin{bmatrix} -0.5462 \\ -0.4232 \\ 0.9130 \end{bmatrix}$$

For instance, if the i^{th} observation contains value 2 for Sex, 1 for $SMOKESTAT$ and 3 for $VITUSE$. Therefore the linear combination of the weights of these categories would yield -0.2661 + -0.6169 + 0.9130 = -0.03. Hence the predicted $\hat{Y_i}$ is given by

$$\begin{aligned} \begin{bmatrix} 1 & -0.03 \end{bmatrix} \times beta &= \begin{bmatrix} 1 & -0.03 \end{bmatrix} \begin{bmatrix} 189.8921 & 602.7905 \\ -29.7136 & 18.4572 \end{bmatrix} \\ &= \begin{bmatrix} 190.7835 & 602.23678 \end{bmatrix} \end{aligned}$$

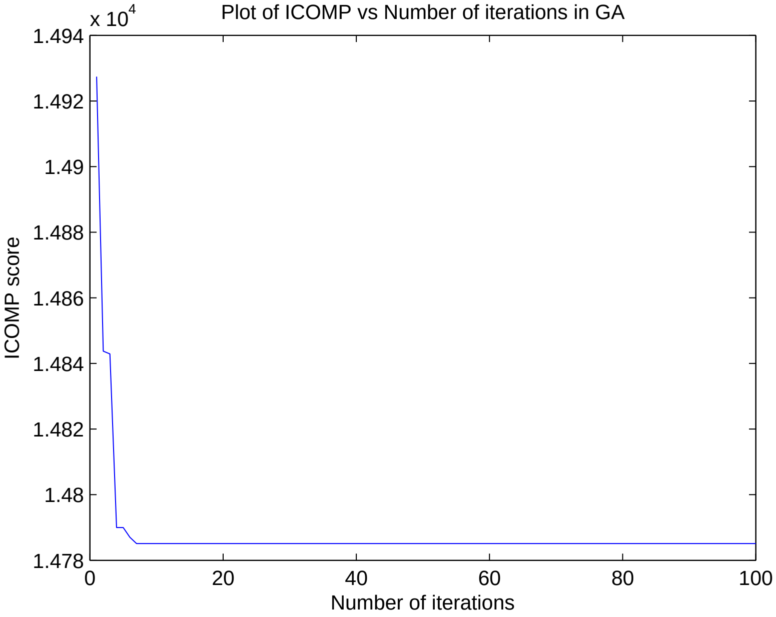


Figure 6.16: Beta-Carotene (MVR): Plot of ICOMP vs Number of iterations in GA

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.16. The set of variables selected by AIC with same set of GA parameters is given by

Model: Intercept, Age, Sex, Smokstat, Quetlet, Vituse, Fiber, Cholestrol, Betadiet

The AIC score for this model is 14752.2643.

The estimated parameters for the model given by AIC are given by

$$\beta = \begin{bmatrix} 225.1809 & 485.7480 \\ 1.1767 & 2.7632 \\ -6.2996 & 0.7885 \\ 4.9647 & -1.3943 \\ -0.1360 & -0.1047 \\ 0.0183 & 0.0005 \\ -26.6503 & 16.3297 \end{bmatrix}$$

Since ICOMP is more consistent in model selection, we consider the model selected by ICOMP as our best fitting model for this data. The variable selection for multivariate regression in NCSS gave Age, Sex, Quetlet, Vituse, Fiber, Cholestrol, Betadiet as the best predictors for this data in the original mixed space. Even in the multivariate regression case, the most obvious categorical variable Smokestat is not picked up by the model in the original mixed data space whereas it is picked up by the model in the Gifi space.

Beta-Carotene – Optimal Scaling Method

The data is transformed to the Gifi space and the categorical variables are optimally scaled. We fit a multivariate regression model for the Beta-Carotene data with BETAPLASMA and RETPLASMA as dependent variables and AGE, SEX, SMOKSTAT, QUETELET, VITUSE, CALORIES, FAT, FIBER, ALCOHOL, CHOLESTROL, BETADIET and RETDIET as independent variables. Since the variables SEX, SMOKSTAT, and VITUSE are categorical, we use the Gifi transformation to come up with an optimal weight vector that is used for transforming the categorical space to a pure continuous space. We assume that the residuals are multivariate normally distributed. We also include the intercept term in this model. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The best set of variables selected by GA and its associated information score is given by

Intercept, Age, Sex, Smokstat, Quetlet, Vituse, Fiber

The information criteria score is: 8381.3187.

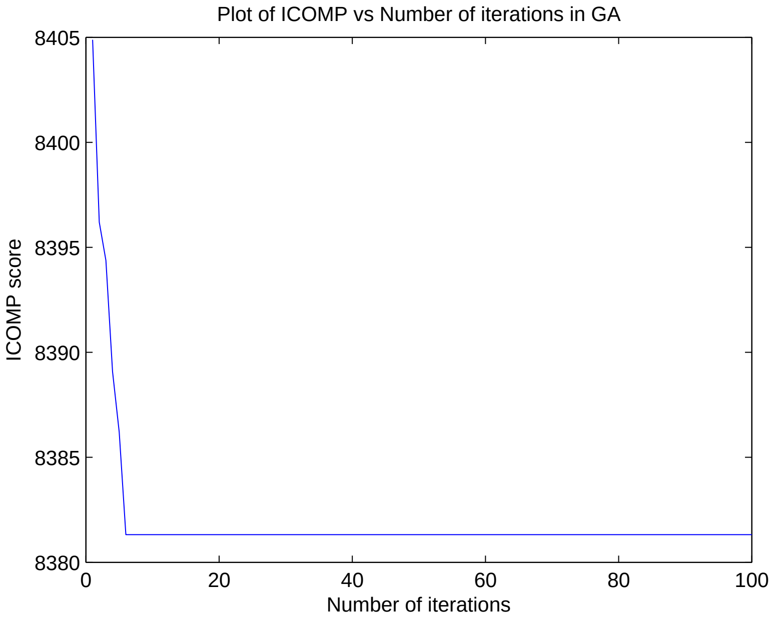


Figure 6.17: Beta-Carotene (MVR-OS): Plot of ICOMP vs Number of iterations in GA

The parameter estimates for the above model is given by

$$\beta = \begin{bmatrix} 195.6098 & 475.9395 \\ 1.4726 & 2.6944 \\ -23.9980 & 42.0694 \\ -18.3848 & 26.1379 \\ -6.3072 & 0.7824 \\ -47.4683 & -24.5633 \\ 6.6793 & -2.2463 \end{bmatrix}$$

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.17. The set of predictors selected by AIC are given by

Intercept, Age, Sex, Quetlet, Vituse, Fiber, Cholestrol, Betadiet

The AIC score for this model: 8381.5021.

Since ICOMP is more consistent in model selection, we choose the model selected by ICOMP as our best fitting model.

6.4.2 Cars Data

Cars – Linear Combination Method

We fit a multivariate regression to the cars data with SRP, CMPG, and HMPG as the dependent variables and the variable Sports Car, SUV, Wagon, Minivan, Pickup, AWD, RWD, NumCylinders, Engine size, HP, Weight, Wheel Base, Length and Width as predictor variables. We assume a multivariate normal distribution on the residuals. We also include the intercept term in this model. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The best set of variables selected by GA and its associated information score is given by

Model: Intercept, SUV, Wagon, Minivan, EngSize, HP

Information criteria score: 26333.2606.

The parameter estimates for the above set of variables is given by

$$\beta = \begin{bmatrix} 1.0e+004* & & \\ -1.6021 & 0.0032 & 0.0039 \\ -0.1915 & -0.0002 & -0.0002 \\ 0.0258 & -0.0000 & -0.0000 \\ -0.1714 & -0.0002 & -0.0004 \end{bmatrix}$$

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.18. The set of variables selected by AIC with the same set of GA parameters is given by

Model: Intercept, SUV, Wagon, Minivan, HP, Weight, WheelBase, Length, Width

The AIC score for this model is 20162.0113.

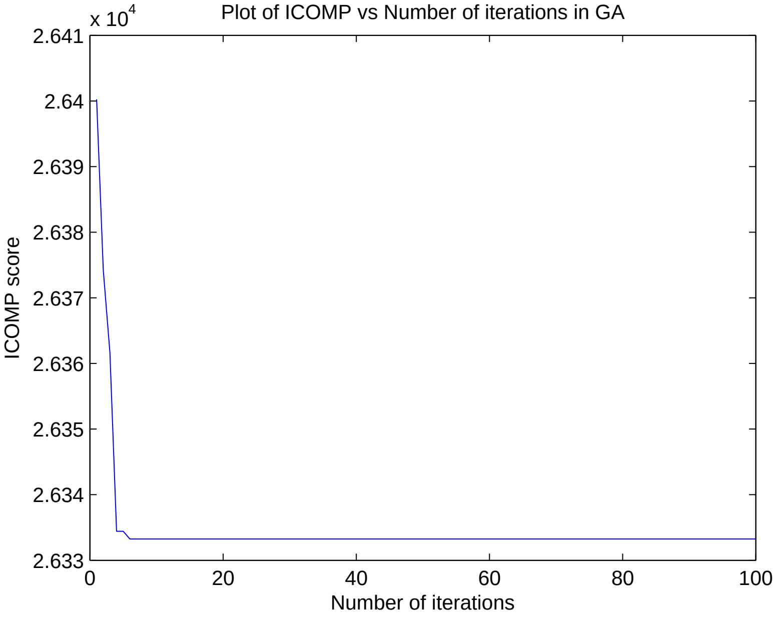


Figure 6.18: Cars Data (MVR): Plot of ICOMP vs Number of iterations in GA

The estimated parameters for the model given by AIC are given by

$$\beta = \begin{bmatrix} 69010 & 34 & 32 \\ 225 & 0 & 0 \\ 9 & 0 & 0 \\ -617 & 0 & 0 \\ -19 & 0 & 0 \\ -636 & 0 & 0 \\ -4855 & 0 & -2 \end{bmatrix}$$

We consider the model selected by ICOMP as our best fitting model for this data. The variable selection for multivariate regression in NCSS gave SUV, RWD, HP, Weight, WheelBase, and Length as the best predictors for this data in the original mixed space.

Cars – Optimal Scaling Method

The data is transformed to the Gifi space and the categorical variables are optimally scaled. We fit a multivariate regression to the cars data with SRP, CMPG, and HMPG as the dependent variables and the variable Sports Car, SUV, Wagon, Minivan, AWD, RWD, NumCylinders, Engine size, HP, Weight, Wheel Base, Length and Width as predictor variables. We assume a multivariate normal distribution on the residuals. We also include the intercept term in this model. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The best set of variables selected by GA and its associated information score is given by

Intercept, SUV, RWD, EngSize, NumCylinders, HP, Weight, WheelBase

The information criteria score : 17660.0141.

The parameter estimates are given by

$$\beta = \begin{bmatrix} 1.0e+004* & & \\ 3.9084 & 0.0032 & 0.0033 \\ -0.3516 & -0.0000 & -0.0002 \\ -0.3184 & 0.0000 & 0.0001 \\ -0.3418 & -0.0001 & -0.0000 \\ -0.7702 & -0.0002 & -0.0001 \\ 0.0213 & -0.0000 & -0.0000 \\ 0.0009 & -0.0000 & -0.0000 \\ -0.0683 & 0.0000 & 0.0000 \end{bmatrix}$$

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.19. The best set of predictors selected by AIC is given by

Intercept, SUV, Minivan, AWD, RWD, NumCylinders, HP, Weight, WheelBase, Length, Width

The information criteria score: 11590.1955.

We choose the model selected by ICOMP as our best fitting model.

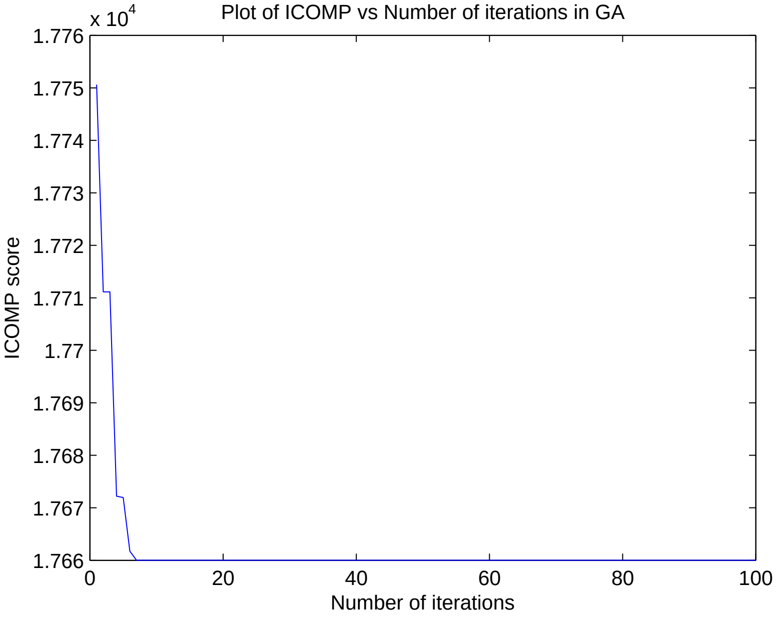


Figure 6.19: Cars Data (MVROS): Plot of ICOMP vs Number of iterations in GA

6.5 Gifi - Multivariate Logistic Regression

We use Gifi transformation on a mixed data set to transform the categorical predictor variables to a pure continuous space and then fit a multivariate logistic regression model to predict a set of categorical responses. We show the results of our algorithm on a real data set.

Healthcare Service Data

Healthcare Service Data – Linear Combination Method

This data contains 98 observations on 9 independent and 2 binary dependent variables. Some of independent variables are continuous and some are categorical. These observations are drawn from a population of more than 9000 cases. This data set is from a health care company that offers home health service to patients, usually old people, and gets their revenue from Medicare payments. The company experienced a great amount of losses. So to improve their operating strategies, the company decided to introduce statistical analysis to build up models to identify the factors (variables) that are important in determining y1, whether this patient is profitable ('1') or not ('0') and y2, whether this patient is hospitalized ('1') or not ('0').

The variables are briefly described below:

- LOS (Length of stay). Type: Continuous
- AGE (Age of a patient). Type: Continuous
- ICD: Code for the primary disease of the patient. Type: Categorical
- REHPOT: Potential for rehabilitation. Type: Categorical
- FREQ: Number of times care giver comes per day. Type: Categorical (1 - more than 10 times, 2 - 8 to 10 times, 3 - 5 to 7 times, 4 - 1 to 4 times, 5 - irregularly, 6 - unknown)
- TOTVS: Total visits from nurses and therapists for an episode. Type: Continuous
- SUPPCHGS: Supply charges. Type: Continuous
- SEX: Gender of a patient. Type: categorical
- GRP: number of a group into which a patient is classified according to different clinical scores, functional scores, and service scores.
- PROFITABLE: Is it profitable to provide services to a particular patient. Type: Categorical.
- HOSPITALIZED: Is a particular patient HOSPITALIZED? Type: Categorical.

A plot matrix of the data in the Gifi space is shown in figure 6.20. The plot matrix is in the order of the predictors LOS, AGE, TOTVS, SUPPCHGS and catX where catX is the linear combination of the categories of the categorical variables in the Gifi space. Some of the pair-wise kernel density estimates of this data in the Gifi space is shown in table 6.5.

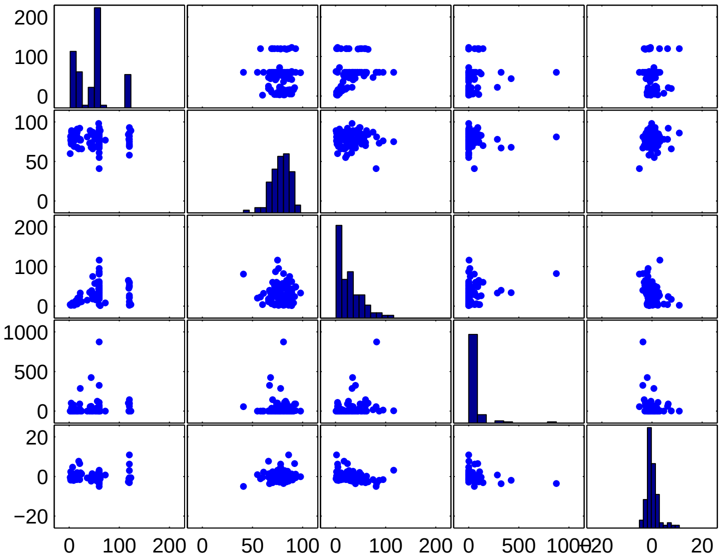
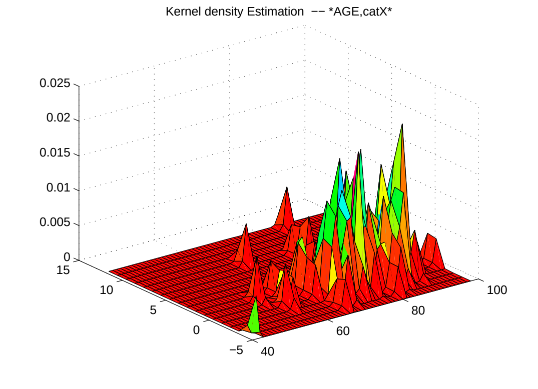
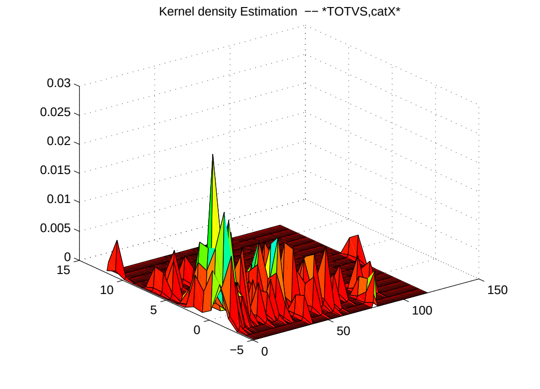


Figure 6.20: HCS Data: Plot Matrix of the data in the Gifi space

Table 6.28: Kernel Density Estimate of Health Serivce Data in the Gifi space

KDE1	KDE2
	

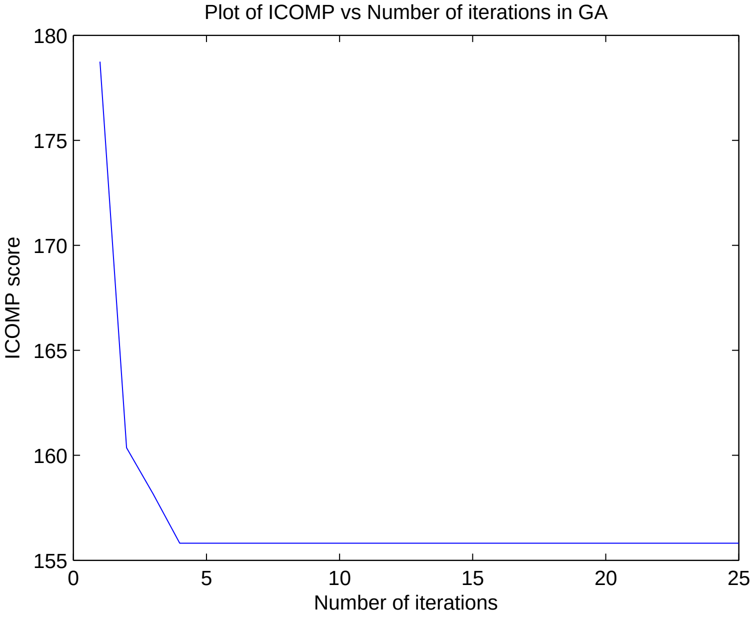


Figure 6.21: HCS Data (MVLR): Plot of ICOMP vs Number of iterations in GA

The categorical predictor variables are first transformed to 1-dimensional Gifi space and a multivariate logistic regression model is fit to the data with PROFITABLE and HOSPITALIZED as binary responses and LOS, AGE, ICD, REHPOT, TOTVS, SUPPCHGS, GRP as predictors. The input model uses the intercept term. We use GA for variable selection with maximum iterations of 20, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{IFIM}$ is used as the fitness function.

The following model is selected by GA process with the above input parameters.

$$\text{Model} = \text{Intercept, ICD, TOTVS, SEX}$$

Information Criterion Score: 155.8194

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.21. The estimated parameters are given by

Table 6.29: HCS: Linear Combination Values

PROFITABLE	HOSPITALIZED	Linear Combination Value (LCV)
0	0	(0.0265 + (-0.4905)) = -0.464
0	1	(0.0265 + (0.4709)) = 0.4974
1	0	(-0.4058 + (-0.4905)) = -0.8963
1	1	(-0.4058 + 0.4709) = 0.0651

$$\beta = \begin{bmatrix} -5.1460 & -1.8936 & -2.4355 \\ 0.0829 & 0.0723 & -0.0181 \\ -0.9179 & -0.7818 & -0.2603 \end{bmatrix}$$

where the first row corresponds to the intercept coefficient and the second row corresponds to the coefficient of the continuous predictor TOTVS, and the third row corresponds to the coefficient of the categorical variables, ICD and SEX.

The weight vectors associated with the categorical response variables, PROFITABLE and HOSPITALIZED, are given by

$$wPROFITABLE = \begin{bmatrix} 0.0265 \\ -0.4058 \end{bmatrix} \qquad wHOSPITALIZED = \begin{bmatrix} -0.4905 \\ 0.4709 \end{bmatrix}$$

Hence, a linear combination of the weights of these two categorical response variables would give four different continuous values given in table 6.29. Hence, the 1-dimensional continuous values in the Gifi space for the categorical response variables PROFITABLE, and HOSPITALIZED consists of four distinct values. We consider these four distinct LCV's as four different classes with the lowest LCV as class 1, the next lowest as class 2, the next lowest as class 3, and the highest LCV as class 4. Hence the linear combination that corresponds to PROFITABLE = 1 and HOSPITALIZED = 0 is considered as class 1, the linear combination corresponding to PROFITABLE = 0 and HOSPITALIZED = 0 is considered as class 2, the linear combination corresponding to PROFITABLE = 1 and HOSPITALIZED = 1 is considered as class 3 and the linear combination corresponding to PROFITABLE = 0 and HOSPITALIZED = 1 is considered as class 4.

Table 6.30: ICOMP: Confusion Matrix (HCS)

Class	1	2	3	4
1	0	0	0	0
2	3	31	0	7
3	0	0	0	0
4	0	14	3	40

Table 6.31: AIC: Confusion Matrix (HCS)

Class	1	2	3	4
1	0	0	0	0
2	3	31	0	9
3	0	0	0	0
4	0	14	3	38

The confusion matrix (ICOMP) is given in table 6.30. The prediction accuracy is 72.45% and the error rate is 27.55%.

The model selected by AIC is given by

$$\text{Model} = \text{Intercept, LOS, ICD, TOTVS, SUPPCHGS, SEX}$$

The AIC score for this model is 137.8224.

The parameter estimates are given

$$\beta = \begin{bmatrix} -5.2812 & -2.1566 & -1.8322 \\ 0.0032 & 0.0056 & -0.0977 \\ 0.0758 & 0.0664 & 0.0860 \\ 0.0094 & 0.0079 & -0.0085 \\ -0.7962 & -0.7728 & -0.1817 \end{bmatrix}$$

The confusion matrix (AIC) is given in table 6.31. The prediction accuracy is 70.41% and the error rate is 29.59%.

We consider the model selected by ICOMP as the best fitting model.

Healthcare Service Data – Optimal Scaling Method

The data set if first transformed to the Gifi space and the categorical variables are optimally scaled. In the Gifi space, the data set is purely continuous. Hence, we fit a multivariate regression with PROFITABLE and HOSPITALIZED variables in the Gifi space as responses and the variables LOS, AGE, ICD, REHPOT, FREQ, TOTVS, SUPPCHGS, SEX, GLEASON variables in the Gifi space as predictors. The input model includes the intercept term. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{IFIM}$ is used as the fitness function.

The following model is selected as best model by the GA process with above input parameters.

Intercept, ICD, REHPOT, FREQ, TOTVS, GRP

The information criteria score: -7.7717.

The estimated parameters for this model is given by

$$\beta = \begin{bmatrix} -0.0077 & 0.2737 \\ 0.0071 & 0.2645 \\ -0.0300 & -0.1525 \\ -0.0038 & -0.1129 \\ 0.0003 & -0.0098 \\ 0.0256 & -0.1307 \end{bmatrix}$$

The predicted values of the response variables PROFITABLE and HOSPITALIZED are re-mapped to their original scale. The confusion matrix for the response variable, PROFITABLE, is given in table 6.32. The prediction accuracy is 93.88% and the error rate is 6.12%.

The confusion matrix for the response variable, HOSPITALIZED, is given in table 6.33. The prediction accuracy is 81.63% and the error rate is 18.37%.

Table 6.32: Classification matrix

PROFITABLE	*0*	*1*
0	92	6
1	0	0

Table 6.33: Classification matrix

HOSPITALIZED	*0*	*1*
0	37	7
1	11	43

The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.22.

6.6 Gifi - Discriminant Analysis

We use Gifi transformation on a mixed data set to transform the categorical predictor variables to a one dimensional continuous space and run discriminant analysis in the Gifi space. We show the results of our algorithm on two real data sets.

6.6.1 ICU Data

ICU Data: Linear Combination Method

The data is run for discriminant analysis in the original space with STA as the classification variable and AGE, SEX, RACE, SER, CAN, CRN, INF, CPR, SYS, HRA, PRE, TYP, FRA, PO2, PH, PCO, BIC, CRE, LOC as predictor variables. The classification variable, STA, has two groups (0 and 1). We use the likelihood ratio statistic to test for the equality of the covariance matrices for the two groups, ([Ender, 1998]). The covariance matrices for the two groups in the original space are detected to be unequal. Hence, we run quadratic discriminant analysis instead of linear discriminant analysis. The confusion matrix is given in table 6.34. The AIC score for this model is 7042.2 and ICOMP score is 7632.9. The prediction accuracy is 87%. The following classification functions are computed for each

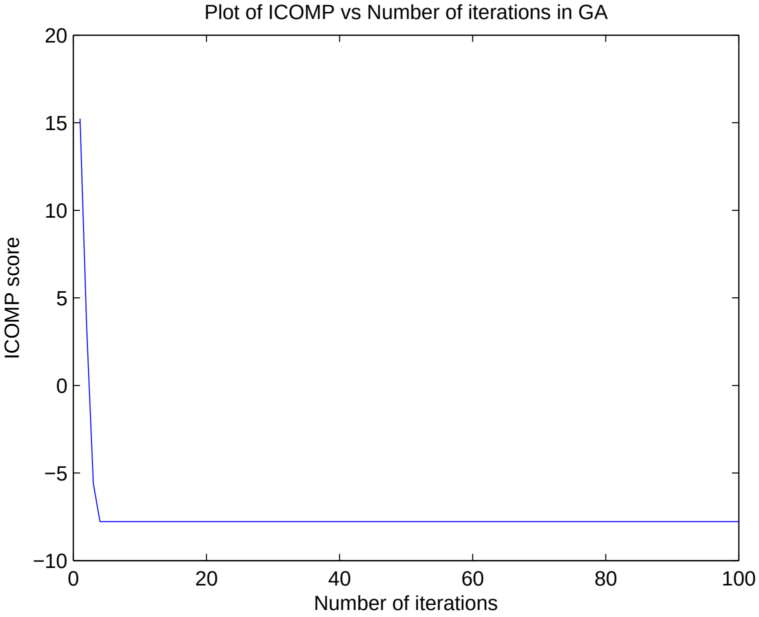


Figure 6.22: HCS Data (MVLROS): Plot of ICOMP vs Number of iterations in GA

Table 6.34: ICU Data: Confusion Matrix

STA	0	1
0	140	6
1	20	34

category.

$$\begin{aligned}
STA(0) \quad = \quad & -171.4033 + 0.1647 \times AGE + 1.9163 \times SEX + \\
& 4.5249 \times RACE + 25.4714 \times SER + 25.2614 \times CAN + \\
& 2.7593 \times CRN + 2.8873 \times INF + 41.4321 \times CPR + \\
& 0.2245 \times SYS + 0.0370 \times HRA + 15.7151 \times PRE + \\
& 23.0880 \times TYP + 18.3646 \times FRA + 32.9669 \times PO2 + \\
& 8.9991 \times PH - 10.0003 \times PCO + 10.4094 \times BIC + \\
& 31.5575 \times CRE + 18.7705 \times LOC
\end{aligned}$$

$$\begin{aligned}
STA(1) \quad = \quad & -444.4892 + 1.2151 \times AGE + -54.5064 \times SEX + \\
& 24.6862 \times RACE + 49.0009 \times SER + 57.0335 \times CAN + \\
& -25.1709 \times CRN - 38.2563 \times INF - 2.5582 \times CPR + \\
& 0.6841 \times SYS + 1.1966 \times HRA + 73.7960 \times PRE + \\
& 150.4803 \times TYP + 49.8615 \times FRA - 4.5000 \times PO2 - \\
& 0.8055 \times PH + 40.9973 \times PCO + 26.8239 \times BIC + \\
& 56.5803 \times CRE + 34.2815 \times LOC
\end{aligned}$$

Table 6.35: ICU Data: Confusion Matrix

STA	0	1
0	147	23
1	13	17

The number of discriminant functions selected by ICOMP is 1. It is given by

$$DF = \begin{bmatrix} -0.0075 \\ 0.0885 \\ 0.0121 \\ 0.0754 \\ -0.4206 \\ -0.0086 \\ -0.0407 \\ -0.1075 \\ 0.0016 \\ 0.0007 \\ -0.1783 \\ -0.3609 \\ -0.1122 \\ -0.0783 \\ -0.3564 \\ 0.3902 \\ 0.0683 \\ -0.0871 \\ -0.5690 \end{bmatrix}$$

Now, we present the results of the discriminant analysis in the Gifi space. The categorical predictors are transformed to the Gifi space. The covariance matrices for the two groups are detected to be unequal. Hence we run quadratic discriminant analysis in the Gifi space. The confusion matrix is shown in table 6.35. The AIC score for this model is 6505.6 and ICOMP score is 6515.1. The prediction accuracy is 82%. There is a 5% loss in the prediction accuracy. This loss can be accounted for the loss of information using the LCM procedure. The following classification functions are computed for each category.

$$STA(0) = -20.0056 + 0.1099 \times AGE + 0.1319 \times SYS + 0.1551 \times HRA + 0.6305 \times catX$$

$$STA(1) = -24.6195 + 0.2121 \times AGE + 0.1281 \times SYS + 0.1635 \times HRA - 0.3037 \times catX$$

where catX is the linear combination of the weights of the corresponding categories of the categorical predictors in the Gifi space. The weight vector associated with SEX, RACE, SER, CAN, CRN, INF, CPR, PRE, TYP, FRA, PO2, PH, PCO, BIC, CRE, LOC are given by

$$wSEX = \begin{bmatrix} 0.1478 \\ -0.2411 \end{bmatrix} \qquad wRACE = \begin{bmatrix} -0.0178 \\ 0.1572 \\ 0.0757 \end{bmatrix} \qquad wSER = \begin{bmatrix} -0.7401 \\ 0.6433 \end{bmatrix}$$

$$wCAN = \begin{bmatrix} -0.0823 \\ 0.7405 \end{bmatrix} \qquad wCRN = \begin{bmatrix} 0.1362 \\ -1.2978 \end{bmatrix} \qquad wINF = \begin{bmatrix} 0.3867 \\ -0.5341 \end{bmatrix}$$

$$wCPR = \begin{bmatrix} 0.1171 \\ -1.6840 \end{bmatrix} \qquad wPRE = \begin{bmatrix} -0.0352 \\ 0.1997 \end{bmatrix} \qquad wTYP = \begin{bmatrix} 0.9797 \\ -0.3532 \end{bmatrix}$$

$$wFRA = \begin{bmatrix} -0.0140 \\ 0.1725 \end{bmatrix} \qquad wP02 = \begin{bmatrix} 0.1172 \\ -1.3476 \end{bmatrix} \qquad wPH = \begin{bmatrix} 0.1349 \\ -1.9410 \end{bmatrix}$$

$$wPCO = \begin{bmatrix} 0.1118 \\ -1.0059 \end{bmatrix} \qquad wBIC = \begin{bmatrix} 0.1376 \\ -1.6966 \end{bmatrix} \qquad wCRE = \begin{bmatrix} 0.0948 \\ -1.8007 \end{bmatrix}$$

$$wPCO = \begin{bmatrix} 0.1087 \\ -0.4921 \\ -1.7652 \end{bmatrix}$$

Suppose if a new observation contains the following values:

AGE = 27

SEX = 1

RACE = 1

SER = 1

CAN = 1

CRN = 1

INF = 1

CPR = 1

SYS = 142

HRA = 88

PRE = 1

TYP = 1

FRA = 1

PO2 = 1

PH = 1

PCO =1

BIC = 1

CRE = 1

LOC = 1

In the Gifi space, the value of catX for this observation is given by

$$\begin{aligned}
catX &= -0.2411 - 0.0178 + 0.6433 + 0.7405 - \\
&\quad 1.2978 - 0.5341 - 1.6840 + 0.1997 - 0.3532 + \\
&\quad 0.1725 - 1.3476 - 1.9410 - 1.0059 - 1.6966 - \\
&\quad 1.8007 - 0.4921 \\
&= -10.6559
\end{aligned}$$

Hence,

$$\begin{aligned}
STA(0) &= -20.0056 + 0.1099 \times 27 + 0.1319 \times 142 + \\
&\quad 0.1551 \times 88 + 0.6305 \times (-10.6559) \\
&= 8.6217
\end{aligned}$$

$$\begin{aligned}
STA(1) &= -24.6195 + 0.2121 \times 27 + 0.1281 \times 142 + \\
&\quad 0.1635 \times 88 - 0.3037 \times (-10.6559) \\
&= 16.9217
\end{aligned}$$

Since $STA(1) > STA(0)$, we assign the new observation to the group for which $STA = 1$.

The within group and between group covariance matrices in the Gifi space are given by

$$\Sigma_W = \begin{bmatrix} 385.8 & 53.5 & 16.8 & -3.1 \\ 53.5 & 1035.2 & -44.1 & 11.4 \\ 16.8 & -44.1 & 715.5 & -22.7 \\ -3.1 & 11.4 & -22.7 & 6.2 \end{bmatrix}$$

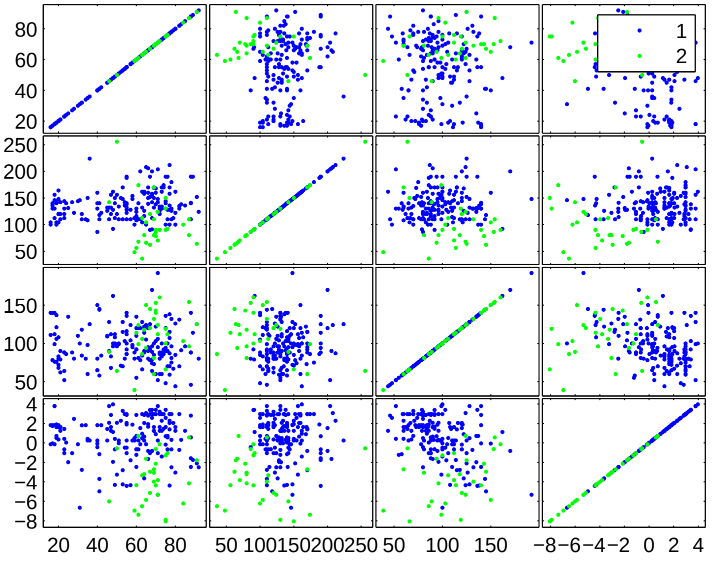


Figure 6.23: ICU Data (DA): Scatter Plot Matrix of the data in the Gifi space

$$\Sigma_B = \begin{bmatrix} 14.3641 & -25.4972 & 3.2215 & -3.3754 \\ -25.4972 & 45.2593 & -5.7184 & 5.9916 \\ 3.2215 & -5.7184 & 0.7225 & -0.7570 \\ -3.3754 & 5.9916 & -0.7570 & 0.7932 \end{bmatrix}$$

This is one of the advantages of the Gifi system since the dimension of the within group and between group covariance matrices would be much less than in the original space if there are many categorical variables in the data.

The number of discriminant functions selected by ICOMP are 2. They are given by

$$\begin{bmatrix} DF1 & DF2 \end{bmatrix} = \begin{bmatrix} -0.2862 & -0.0677 \\ -0.0379 & 0.0394 \\ -0.0263 & 0.0273 \\ -0.9570 & 0.9966 \end{bmatrix}$$

The scatter plot matrix for each pair of variables in the Gifi space is shown in figure 6.23. The set of variables that are included in the scatter plot are AGE, SYS, HRA and catX

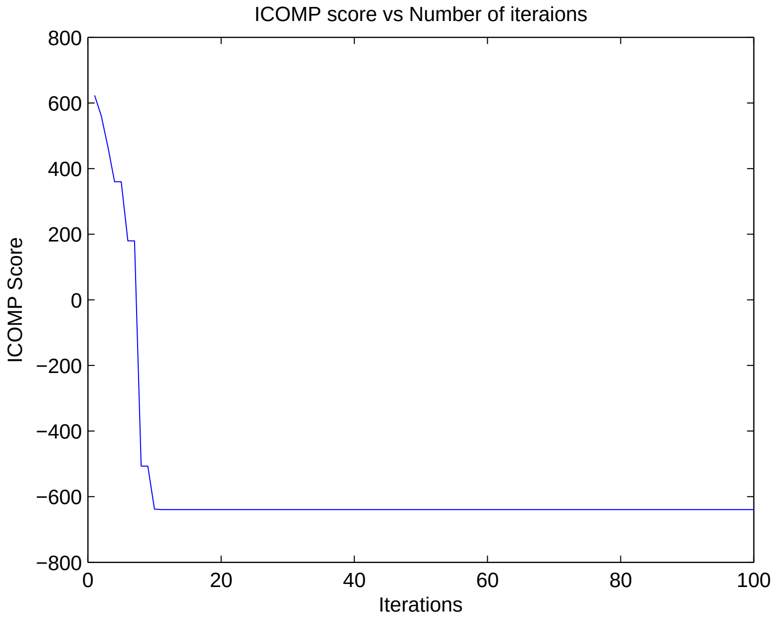


Figure 6.24: ICU (DA): Plot of ICOMP vs Number of iterations in GA

where catX is the linear combination of the categories of the categorical variables in the Gifi space. The scatterplot on the first row and the second column is the scatterplot of the variable AGE and SYS. The two groups are identified with two different colors in the scatterplot.

Now, we show the results of the variable selection on this data set. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1F}$ is used as the fitness function.

The following parameters are selected by GA with the above input parameters.

$$\text{Model : RAC}$$

Information Criterion Score: -639.0167. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.24.

Table 6.36: ICU Data: Confusion Matrix

STA	0	1
0	144	8
1	16	32

ICU Data: Optimal Scaling Method

The data is transformed to the Gifi space and the categorical variables are optimally scaled. We fit a discriminant analysis with STA as the classification variable and AGE, SEX, RACE, SER, CAN, CRN, INF, CPR, SYS, HRA, PRE, TYP, FRA, PO2, PH, PCO, BIC, CRE, LOC as predictor variables. The classification variable, STA, has two groups (0 and 1). The covariance matrices for the two groups in the Gifi space are detected to be unequal. Hence, we run quadratic discriminant analysis instead of linear discriminant analysis. The confusion matrix is given in table 6.36. The AIC score for this model is 5954 and ICOMP score is 7485.2. The prediction accuracy is 88% and the error rate is 12%.

The classification functions are given by

$$\begin{aligned} STA(0) &= -24.6708 + 0.1898 \times AGE + 1.3122 \times SEX \\ &\quad +12.3108 \times RAC + 1.1111 \times SER + 0.8713 \times CAN \\ &\quad +5.9038 \times CRN + 2.0444 \times INF - 1.6979 \times CPR \\ &\quad +0.1567 \times SYS + 0.1694 \times HRA - 10.4174 \times PRE \\ &\quad -1.6696 \times TYP + 26.5411 \times FRA + 3.0161 \times PO2 \\ &\quad +0.9725 \times PH + 2.4430 \times PCO - 4.1856 \times BIC \\ &\quad +4.6344 \times CRE + 0.3172 \times LOC \end{aligned}$$

$$
\begin{aligned}
STA(1) \quad = \quad & -78.6236 + 0.6079 \times AGE + 51.0637 \times SEX \\
& +10.2779 \times RAC + 9.3886 \times SER + 11.9869 \times CAN \\
& +2.5168 \times CRN + 7.9870 \times INF + 6.5007 \times CPR \\
& +0.3757 \times SYS + 0.5917 \times HRA + 91.4165 \times PRE \\
& -34.3489 \times TYP + 128.4370 \times FRA + 3.4783 \times PO2 \\
& -0.6055 \times PH + 1.5200 \times PCO + 2.0856 \times BIC \\
& -6.8098 \times CRE - 15.1322 \times LOC
\end{aligned}
$$

The number of discriminant functions selected by ICOMP is 1. It is given by

$$
DF = \begin{bmatrix} -0.0049 \\ -0.1547 \\ 0.4404 \\ 0.0384 \\ -0.3389 \\ 0.0132 \\ 0.0313 \\ 0.0472 \\ 0.0010 \\ 0.0006 \\ -0.5360 \\ 0.1851 \\ -0.3572 \\ 0.0418 \\ 0.1046 \\ -0.2289 \\ -0.0273 \\ 0.0346 \\ 0.3823 \end{bmatrix}
$$

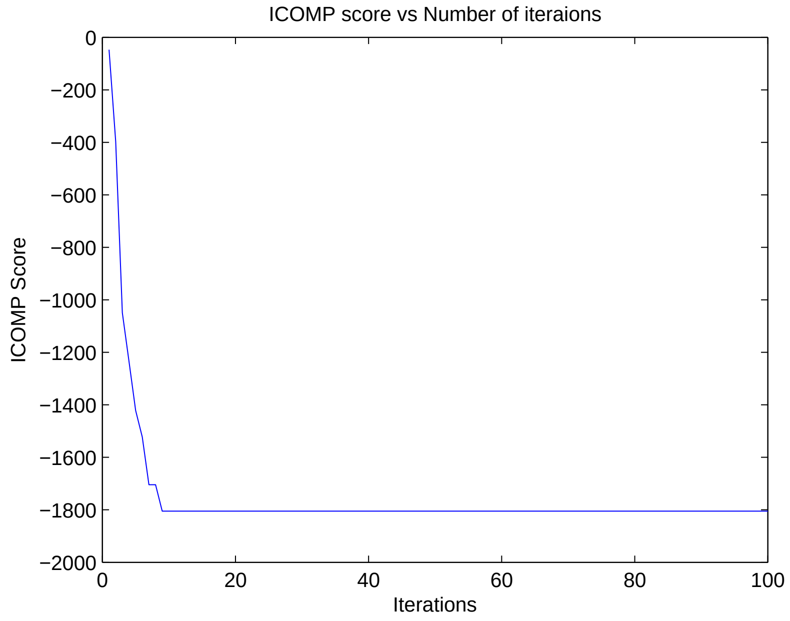


Figure 6.25: ICU (DAOS): Plot of ICOMP vs Number of iterations in GA

Since there are many variables in the Gifi space, we do not give the scatter plot matrix in this case.

Now, we show the results of the variable selection on this data set. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1F}$ is used as the fitness function.

The following parameters are selected by GA with the above input parameters.

Model : SEX, RAC, PRE, FRA

Information Criterion Score: -1805.0682. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.25.

Table 6.37: PCDData: Confusion Matrix

CAPSULE	0	1
0	101	12
1	124	139

6.6.2 Prostate Cancer Data

Prostate Cancer Data: Linear Combination Method

The data is run for discriminant analysis with CAPSULE as the classification variable and AGE, RACE, DPROS, DCAPS, PSA, VOL, GLEASON as predictor variables. The classification variable, CAPSULE, has two groups (0 and 1). The covariance matrices for the two groups in the original space are detected to be unequal. Hence, we run quadratic discriminant analysis instead of linear discriminant analysis. The confusion matrix is given in table 6.37. The AIC score for this model is 11275 and ICOMP score is 11321. The prediction accuracy is 53.72%. The following classification functions are computed for each category.

$$\begin{aligned} CAPSULE(0) &= -85.8067 + 1.5868 \times AGE + 10.9557 \times RACE + \\ & 1.9695 \times DPROS + 22.4489 \times DCAPS - 0.2655 \times PSA + \\ & 0.0083 \times VOL + 4.7656 \times GLEASON \end{aligned}$$

$$\begin{aligned} CAPSULE(1) &= -100.2382 + 1.7384 \times AGE + 18.1691 \times RACE + \\ & 1.5893 \times DPROS + 0.4284 \times DCAPS - 0.1139 \times PSA + \\ & -0.0573 \times VOL + 9.0504 \times GLEASON \end{aligned}$$

The within group and between group covariance matrices in the original space are given by

$$\Sigma_W = \begin{bmatrix} 41.1845 & -0.0860 & -0.2446 & 0.0347 & 0.8503 & 12.9017 & 0.3120 \\ -0.0860 & 0.0866 & 0.0256 & 0.0060 & 0.9123 & 0.4087 & 0.0098 \\ -0.2446 & 0.0256 & 0.8938 & 0.0512 & 2.3555 & -0.3176 & 0.1197 \\ 0.0347 & 0.0060 & 0.0512 & 0.0893 & 1.2666 & -0.4610 & 0.0579 \\ 0.8503 & 0.9123 & 2.3555 & 1.2666 & 353.3217 & 20.4006 & 5.2030 \\ 12.9017 & 0.4087 & -0.3176 & -0.4610 & 20.4006 & 333.4824 & -0.1524 \\ 0.3120 & 0.0098 & 0.1197 & 0.0579 & 5.2030 & -0.1524 & 0.9491 \end{bmatrix}$$

$$\Sigma_B = \begin{bmatrix} 0.0673 & 0.0006 & -0.0823 & -0.0196 & -1.6667 & 0.5463 & -0.1270 \\ 0.0006 & 0.0000 & -0.0008 & -0.0002 & -0.0159 & 0.0052 & -0.0012 \\ -0.0823 & -0.0008 & 0.1005 & 0.0240 & 2.0362 & -0.6674 & 0.1551 \\ -0.0196 & -0.0002 & 0.0240 & 0.0057 & 0.4856 & -0.1592 & 0.0370 \\ -1.6667 & -0.0159 & 2.0362 & 0.4856 & 41.2468 & -13.5189 & 3.1418 \\ 0.5463 & 0.0052 & -0.6674 & -0.1592 & -13.5189 & 4.4309 & -1.0297 \\ -0.1270 & -0.0012 & 0.1551 & 0.0370 & 3.1418 & -1.0297 & 0.2393 \end{bmatrix}$$

The number of discriminant functions selected by ICOMP is 1. It is given by

$$DF = \begin{bmatrix} 0.0116 \\ 0.4231 \\ -0.4206 \\ -0.4434 \\ -0.0168 \\ 0.0090 \\ -0.6686 \end{bmatrix}$$

Now, we present the results of the discriminant analysis in the Gifi space. The categorical predictors are transformed to the Gifi space. The covariance matrices for the two groups are detected to be unequal. Hence we run quadratic discriminant analysis in the Gifi space. The confusion matrix is shown in table 6.38. The AIC score for this model is 10191 and

Table 6.38: ICU Data: Confusion Matrix

STA	0	1
0	48	9
1	177	142

ICOMP score is 10214. The prediction accuracy is 50.53%. There is almost 3% loss in the prediction accuracy. This loss can be accounted for the loss of information using the LCM procedure. The following classification functions are computed for each category.

$$CAPSULE(0) = -55.9138 + 1.6538 \times AGE + 0.1105 \times PSA - 0.0384 \times VOL + 1.3265 \times catX$$

$$CAPSULE(1) = -52.6554 + 1.5595 \times AGE + 0.0362 \times PSA + 0.0124 \times VOL - 0.0467 \times catX$$

where catX is the linear combination of the weights of the categories of the categorical predictors (RACE, DPROS, DCAPS, GLEASON) in the Gifi space.

The within group and between group covariance matrices in the Gifi space are given by

$$\Sigma_W = \begin{bmatrix} 41.1845 & 0.8503 & 12.9017 & -0.2637 \\ 0.8503 & 353.3217 & 20.4006 & -9.1931 \\ 12.9017 & 20.4006 & 333.4824 & 1.2836 \\ -0.2637 & -9.1931 & 1.2836 & 1.6821 \end{bmatrix}$$

$$\Sigma_B = \begin{bmatrix} 0.0673 & -1.6667 & 0.5463 & 0.1832 \\ -1.6667 & 41.2468 & -13.5189 & -4.5338 \\ 0.5463 & -13.5189 & 4.4309 & 1.4860 \\ 0.1832 & -4.5338 & 1.4860 & 0.4984 \end{bmatrix}$$

The number of discriminant functions selected by ICOMP is 1. It is given by

$$DF = \begin{bmatrix} 0.0196 \\ -0.0242 \\ 0.0140 \\ 0.9994 \end{bmatrix}$$

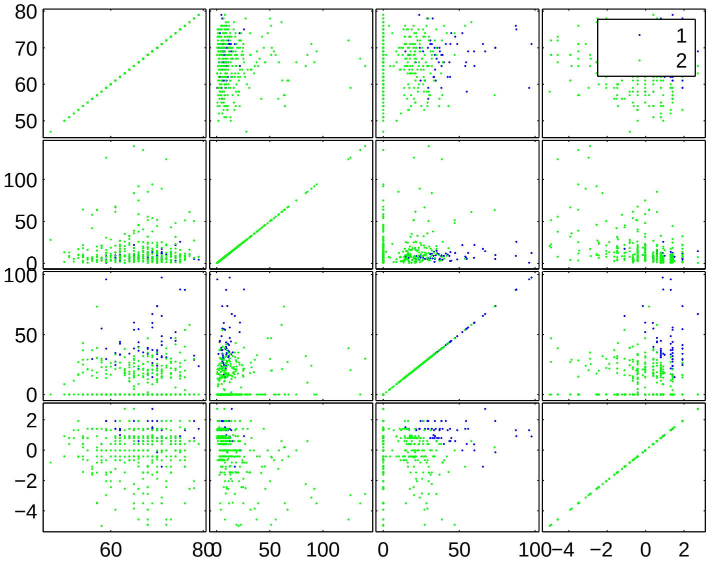


Figure 6.26: PCD Data (DA): Scatter Plot Matrix of the data in the Gifi space

The scatter plot matrix for each pair of variables in the Gifi space is shown in figure 6.26. The set of variables that are included in the scatter plot are AGE, SYS, HRA and catX where catX is the linear combination of the categories of the categorical variables in the Gifi space. The scatterplot on the first row and the second column is the scatterplot of the variable AGE and SYS. The two groups are identified with two different colors in the scatterplot.

Now, we show the results of the variable selection on this data set. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1F}$ is used as the fitness function.

The following parameters are selected by GA with the above input parameters.

$$\text{Model : RACE}$$

Information Criterion Score: -513.1463. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.25.

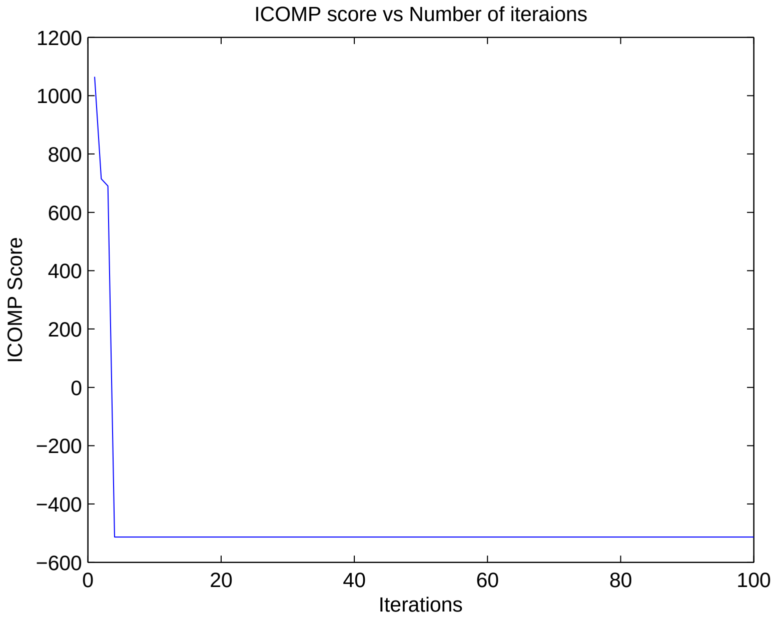


Figure 6.27: PCD (DA): Plot of ICOMP vs Number of iterations in GA

Table 6.39: PCDDataOS: Confusion Matrix

CAPSULE	0	1
0	101	12
1	124	139

Prostate Cancer Data: Optimal Scaling Method

The data is run for discriminant analysis with CAPSULE as the classification variable and AGE, RACE, DPROS, DCAPS, PSA, VOL, GLEASON as predictor variables. The classification variable, CAPSULE, has two groups (0 and 1). The covariance matrices for the two groups in the Gifi space are detected to be unequal. Hence, we run quadratic discriminant analysis instead of linear discriminant analysis. The confusion matrix is given in table 6.39. The AIC score for this model is 10511 and ICOMP score is 10639. The prediction accuracy is 63.829%. The following classification functions are computed for each category.

$$\begin{aligned}
CAPSULE(0) &= -58.0564 + 1.7102 \times AGE + 5.7108 \times RACE + \\
&\quad -1.2613 \times DPROS + 0.7366 \times DCAPS + 0.1478 \times PSA + \\
&\quad -0.0486 \times VOL + 4.2325 \times GLEASON
\end{aligned}$$

$$\begin{aligned}
CAPSULE(1) &= -53.9645 + 1.6063 \times AGE - 12.8706 \times RACE + \\
&\quad 0.6330 \times DPROS + 0.5681 \times DCAPS + 0.0287 \times PSA + \\
&\quad -0.0117 \times VOL - 0.6155 \times GLEASON
\end{aligned}$$

The within group and between group covariance matrices in the original space are given by

$$\Sigma_W = \begin{bmatrix}
41.1845 & 0.0358 & 0.0580 & -0.0703 & 0.8503 & 12.9017 & -0.2872 \\
0.0358 & 0.0150 & 0.0078 & 0.0050 & -0.3792 & -0.1699 & 0.0028 \\
0.0580 & 0.0078 & 0.3724 & 0.0753 & -2.1171 & 0.2202 & 0.0592 \\
-0.0703 & 0.0050 & 0.0753 & 0.3671 & -2.5674 & 0.9345 & 0.0905 \\
0.8503 & -0.3792 & -2.1171 & -2.5674 & 353.3217 & 20.4006 & -4.1294 \\
12.9017 & -0.1699 & 0.2202 & 0.9345 & 20.4006 & 333.4824 & 0.2988 \\
-0.2872 & 0.0028 & 0.0592 & 0.0905 & -4.1294 & 0.2988 & 0.4465
\end{bmatrix}$$

$$\Sigma_B = \begin{bmatrix}
0.0673 & -0.0003 & 0.0510 & 0.0398 & -1.6667 & 0.5463 & 0.0927 \\
-0.0003 & 0.0000 & -0.0002 & -0.0002 & 0.0066 & -0.0022 & -0.0004 \\
0.0510 & -0.0002 & 0.0387 & 0.0301 & -1.2633 & 0.4141 & 0.0702 \\
0.0398 & -0.0002 & 0.0301 & 0.0235 & -0.9843 & 0.3226 & 0.0547 \\
-1.6667 & 0.0066 & -1.2633 & -0.9843 & 41.2468 & -13.5189 & -2.2929 \\
0.5463 & -0.0022 & 0.4141 & 0.3226 & -13.5189 & 4.4309 & 0.7515 \\
0.0927 & -0.0004 & 0.0702 & 0.0547 & -2.2929 & 0.7515 & 0.1275
\end{bmatrix}$$

The number of discriminant functions selected by ICOMP is 1. It is given by

$$DF = \begin{bmatrix} -0.0100 \\ 0.6215 \\ -0.3718 \\ -0.1244 \\ 0.0087 \\ -0.0051 \\ -0.6781 \end{bmatrix}$$

Now, we show the results of the variable selection on this data set. We use GA for variable selection with maximum iterations of 100, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1F}$ is used as the fitness function.

The following parameters are selected by the GA process with above input parameters.

Model : RACE

Information Criterion Score: -513.1463. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.28. In both ICU and prostate cancer data sets, the prediction accuracies in the Gifi space are far better using the discriminant analysis approach than the binary logistic regression approach. This can be accounted for the normalization of the data in the Gifi space. Generally, we recommend the discriminant analysis approach in the case of categorical predictions.

6.7 Gifi - Unsupervised Clustering

6.7.1 ICU Data

ICU Data: Linear Combination Method

The gaussian mixture algorithm is run on the ICU data in the Gifi space using the Linear Combination Method (LCM). The information criteria scores are reported in table 6.40.

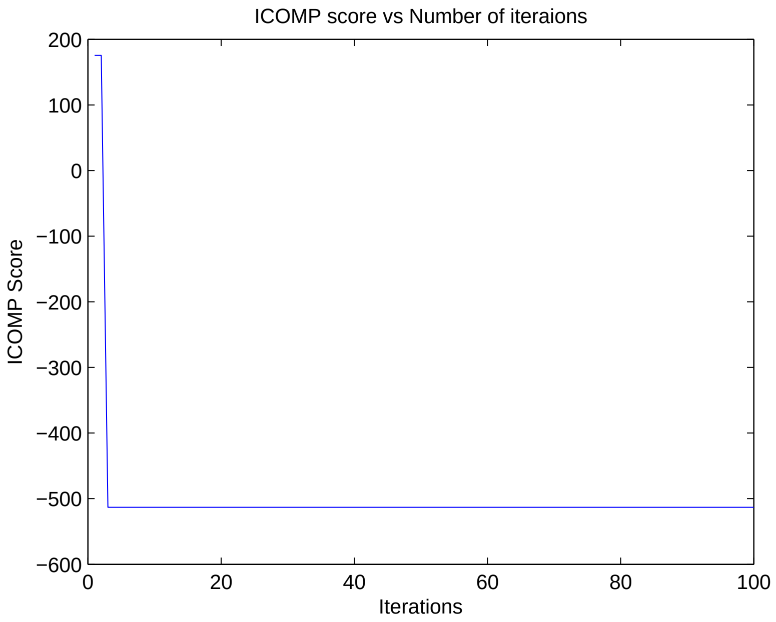


Figure 6.28: PCD (DAOS): Plot of ICOMP vs Number of iterations in GA

Table 6.40: ICU Data: Information Criteria Scores

Mixture	AIC	ICOMP	SBC	CAIC
K = 1	6577.6	6578.3	6609.8	6623.8
K = 2	6517.6	6504.0	**6584.3**	**6613.3**
K = 3	6498.8	6467.6	6599.9	6643.9
K = 4	**6487.5**	**6442.9**	6623.1	6682.1
K = 5	6519.3	6447.8	6689.4	6763.4
K = 6	6541.5	6474.9	6746.1	6835.1
K = 7	6561.2	6476.7	6800.3	6904.3
K = 8	6577.6	6458.6	6851.1	6970.1

Table 6.41: ICU Data: Mixing Proportion Estimates

Mixture	Probability
π_1	0.4018
π_2	0.0902
π_3	0.1362
π_4	0.3719

Table 6.42: ICU Data: Mixture Mean Vector

Mixture	Mean Vector			
K = 1	66.7977	136.8229	110.0631	-0.6007
K = 2	58.9648	119.8873	115.0995	-5.5450
K = 3	19.9852	127.9838	94.8006	1.0346
K = 4	60.9592	131.9496	84.4788	1.6146

AIC and ICOMP are minimum for $K = 4$, where K is the number of gaussian mixtures fitted to the data. SBC and CAIC are minimum for $K = 2$. Since our selection criteria is ICOMP, we consider $K = 4$, as the optimal number of gaussian mixtures fitted to the ICU data.

The mixing proportion estimates are given in table 6.41. The mean vectors for each mixture is given in table 6.42. The covariance matrices for each group are given by

$$\Sigma_1 = \begin{bmatrix} 167.8 & -43.3 & -8.9 & -3.7 \\ -43.3 & 1764.6 & -294.8 & 31.8 \\ -8.9 & -294.8 & 681.5 & -7.8 \\ -3.7 & 31.8 & -7.8 & 4.9 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 201.9 & -63.4 & 2.7 & -7.4 \\ -63.4 & 1487.4 & 390.3 & -3.3 \\ 2.7 & 390.3 & 1090.1 & 25.3 \\ -7.4 & -3.3 & 25.3 & 2.1 \end{bmatrix}$$

$$\Sigma_3 = \begin{bmatrix} 6.9745 & 7.5461 & -23.4536 & -1.8025 \\ 7.5461 & 307.3645 & -71.2192 & 0.2516 \\ -23.4536 & -71.2192 & 659.0682 & -1.4995 \\ -1.8025 & 0.2516 & -1.4995 & 1.6636 \end{bmatrix}$$

Table 6.43: ICU Data: Classification Table

Value	Count	Percent
K = 1	69	34.50%
K = 2	20	10.00%
K = 3	28	14.00%
K = 4	83	41.50%

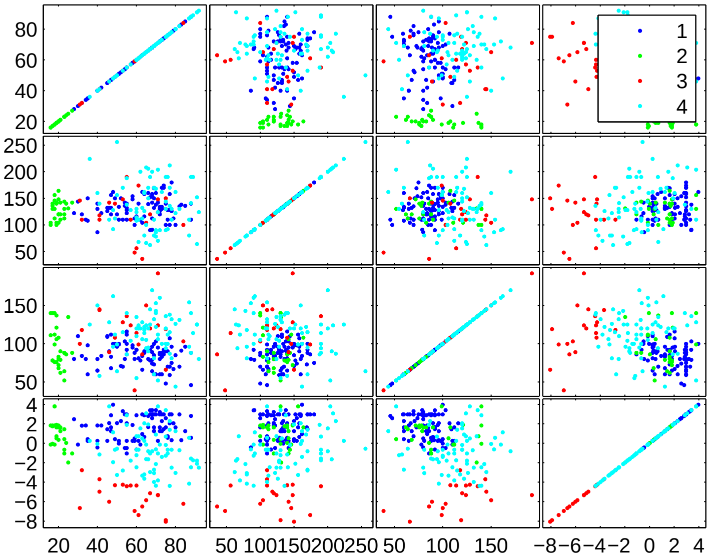


Figure 6.29: ICU Data: Scatter Plot Matrix

$$\Sigma_4 = \begin{bmatrix} 222.0275 & 35.6037 & -52.9731 & 5.5851 \\ 35.6037 & 459.1244 & 98.8143 & 1.5207 \\ -52.9731 & 98.8143 & 271.6618 & -6.4832 \\ 5.5851 & 1.5207 & -6.4832 & 1.6459 \end{bmatrix}$$

The classification table is given in table 6.43. The scatter plot matrix for each pair of variables in the Gifi space is shown in figure 6.29. The set of variables that are included in the scatter plot are AGE, SYS, HRA and catX where catX is the linear combination of the categories of the categorical variables in the Gifi space. The scatterplot on the first row and the second column is the scatterplot of the variable AGE and SYS. The clusters are identified with four different colors in the scatterplot.

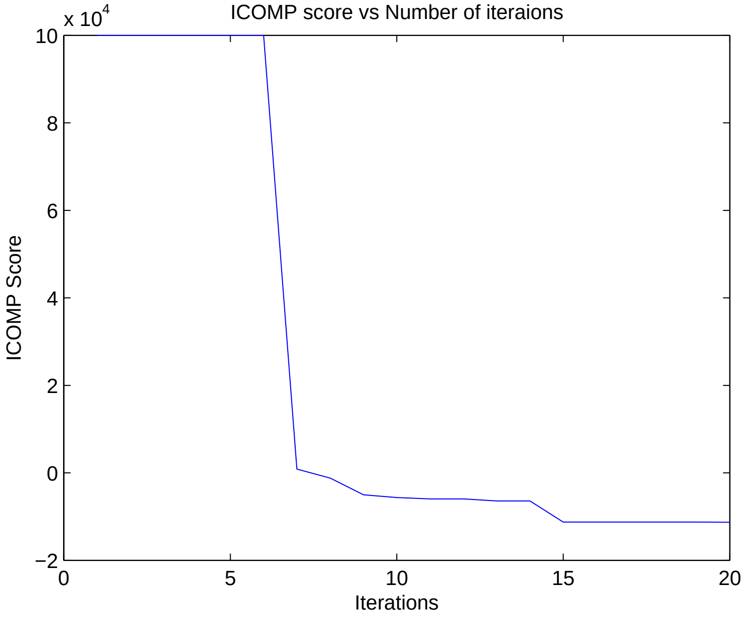


Figure 6.30: ICU (CA): Plot of ICOMP vs Number of iterations in GA

Now, we do a variable selection assuming that the data is generated from four mixtures (since ICOMP is minimum for four mixtures). We use GA for variable selection with maximum iterations of 20, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The following parameters are selected by the GA process with above input parameters.

Model : CPR, TYP, PO2, PH

Information Criterion Score: -11309.7517. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.30. In the original mixed data space, we fit four gaussian mixtures and found that the number of optimal mixtures is 1. The ICU data fits one mixture in the original mixed data space and four mixtures in the Gifi space.

Table 6.44: ICU Data: Information Criteria Scores

Mixture	AIC	ICOMP	SBC	CAIC
K = 1	10335	10339	10376	10390
K = 2	**9805**	**9807**	**9890**	**9919**

Table 6.45: PCD Data: Mixing Proportion Estimates

Mixture	Probability
π_1	0.2388
π_2	0.7612

6.7.2 Prostate Cancer Data

Prostate Cancer Data: Linear Combination Method

The gaussian mixture algorithm is run on the PCD data in the Gifi space using the Linear Combination Method (LCM). The information criteria scores are reported in table 6.44. The algorithm couldn't run for more than 2 mixtures due to the insufficient number of observations in one of the mixtures. Hence the results up to two mixtures are reported for this data. The mixing proportion estimates are given in table 6.45. The mean vectors for each mixture is given in table 6.46. The covariance matrices for each group are given by

$$\Sigma_1 = \begin{bmatrix} 35.2791 & -12.4333 & 13.2839 & -0.7619 \\ -12.4333 & 883.2839 & 94.9101 & -8.5513 \\ 13.2839 & 94.9101 & 253.0473 & -0.0344 \\ -0.7619 & -8.5513 & -0.0344 & 3.3994 \end{bmatrix}$$

$$\Sigma_2 = \begin{bmatrix} 43.0893 & 0.0398 & 13.9949 & 0.3165 \\ 0.0398 & 25.9026 & 17.5486 & -1.1899 \\ 13.9949 & 17.5486 & 357.7300 & 1.1323 \\ 0.3165 & -1.1899 & 1.1323 & 0.8671 \end{bmatrix}$$

Table 6.46: PCD Data: Mixture Mean Vector

Mixture	Mean Vector			
K = 1	66.3544	38.1410	11.8215	-1.5032
K = 2	65.9657	8.1084	17.1595	0.4715

Table 6.47: PCD Data: Classification Table

Value	Count	Percent
K = 1	81	21.54%
K = 2	295	78.46%

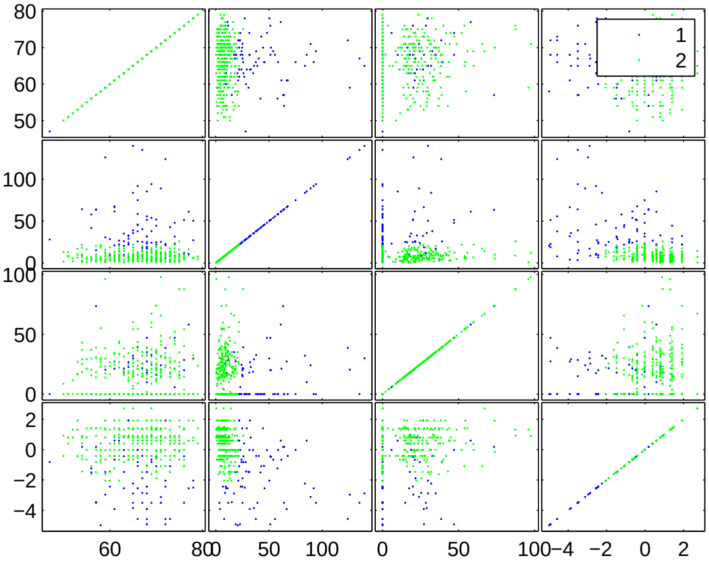


Figure 6.31: PCD Data: Scatter Plot Matrix

The classification table is given in table 6.47. The scatter plot matrix for each pair of variables in the Gifi space is shown in figure 6.31. The set of variables that are included in the scatter plot are AGE, PSA, VOL and catX where catX is the linear combination of the categories of the categorical variables in the Gifi space. The scatterplot on the first row and the second column is the scatterplot of the variable AGE and PSA. The clusters are identified with two different colors in the scatterplot.

Now, we do a variable selection assuming that the data is generated from two mixtures (since ICOMP is minimum for two mixtures). We use GA for variable selection with maximum iterations of 20, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The following parameters are selected by the GA process with above input parameters.

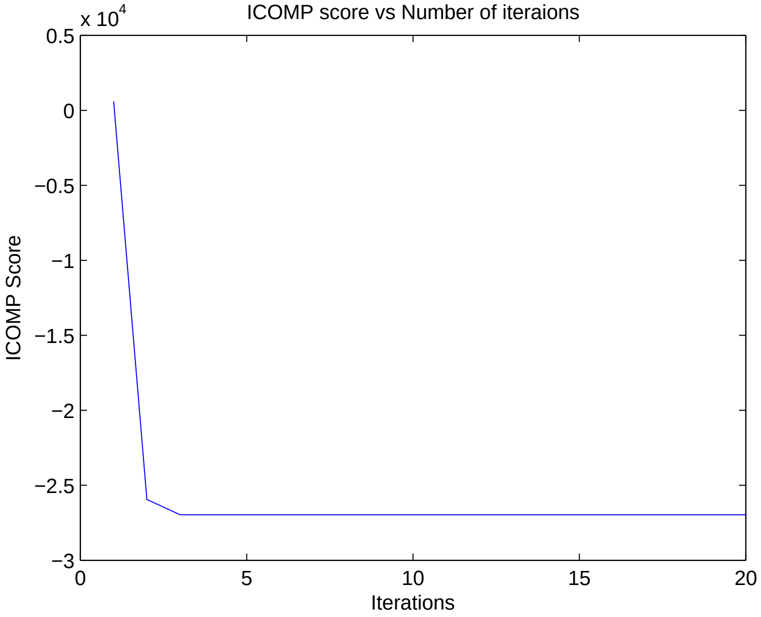


Figure 6.32: PCD (CA): Plot of ICOMP vs Number of iterations in GA

Table 6.48: KCSG1 Data: Information Criteria Scores

Mixture	AIC	ICOMP	SBC	CAIC
K = 1	423.5765	418.3728	426.2641	428.2641
K = 2	**366.2016**	**351.6431**	**372.9206**	**377.9206**

Model : DCAPS

Information Criterion Score: -26966.7369. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.32. In the original mixed data space, we could'nt fit more than two mixtures due to insufficient number of observations in one of the mixtures. This data fits one mixture in the original mixed data space and two mixtures in the Gifi space.

6.7.3 KCS Group 1 Data

KCS Group 1 Data: Linear Combination Method

The gaussian mixture algorithm is run on the KCS group 1 data in the Gifi space using the Linear Combination Method (LCM). The information criteria scores are reported in table 6.48. The algorithm couldn't run for more than 2 mixtures due to insufficient number of

Table 6.49: KCSG1 Data: Mixing Proportion Estimates

Mixture	Probability
π_1	0.4828
π_2	0.5172

Table 6.50: KCSG1 Data: Mixture Mean Vector

Mixture	Mean Vector
K = 1	-3.4764
K = 2	3.2451

observations in one of the mixtures. Hence the results up to two mixtures are reported for this data. The mixing proportion estimates are given in table 6.49. The mean vectors for each mixture is given in table 6.50. The variances for each group are given by

$$\sigma_1 = 3.3956$$

$$\sigma_2 = 0.6675$$

The classification table is given in table 6.51. The scatter plot matrix for each pair of variables in the Gifi space is shown in figure 6.33. The original data contained 40 patients from the KCS group and 37 patients from the non-KCS group. The algorithm was able to identify two mixtures in the Gifi space where one of the mixture contained 37 patients and the other mixture contained 40 patients. Clearly, one mixture indicates the KCS group and the other mixture indicates the non-KCS group. We do not have any information regarding the prior classification of the observations and hence we are unable to provide the confusion matrix for this data set.

Now, we do a variable selection assuming that the data is generated from two mixtures (since ICOMP is minimum for two mixtures). We use GA for variable selection with max-

Table 6.51: KCSG1 Data: Classification Table

Value	Count	Percent
K = 1	37	48.05%
K = 2	40	51.95%

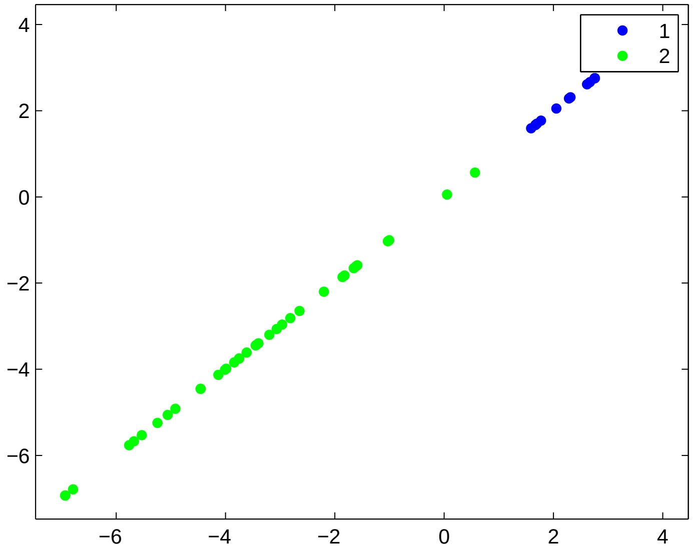


Figure 6.33: KCS Group 1 Data: Scatter Plot Matrix

imum iterations of 20, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The following parameters are selected by GA process with the above input parameters.

$$\text{Model}: \text{G}$$

Information Criterion Score: -5307.5883. The best value of the above fitness function at the end of each iteration of the GA process is shown in figure 6.34. In the original mixed data space, we could'nt fit more than two mixtures due to insufficient number of observations in one of the mixtures. This data fits one mixture in the original mixed data space and two mixtures in the Gifi space.

6.7.4 KCS Group 2 Data

KCS Group 2 Data: Linear Combination Method

The gaussian mixture algorithm is run on the KCS group 2 data in the Gifi space using the Linear Combination Method (LCM). The information criteria scores are reported in table

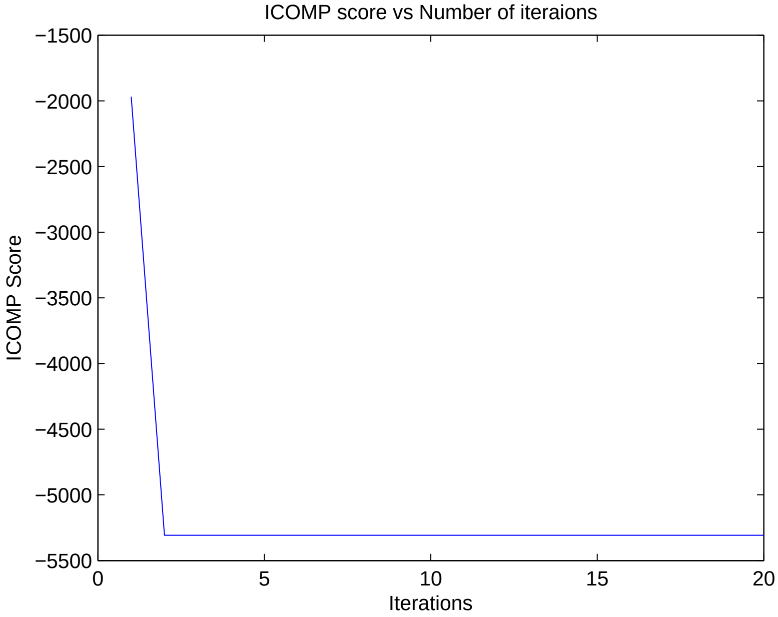


Figure 6.34: KCSG1 (CA): Plot of ICOMP vs Number of iterations in GA

Table 6.52: KCSG2 Data: Information Criteria Scores

Mixture	AIC	ICOMP	SBC	CAIC
K = 1	223.1712	217.8840	224.5983	226.5983
K = 2	**210.0398**	**195.3100**	**213.6077**	**218.6077**

Table 6.53: KCSG2 Data: Mixing Proportion Estimates

Mixture	Probability
π_1	0.3304
π_2	0.6696

Table 6.54: KCSG2 Data: Mixture Mean Vector

Mixture	Mean Vector
K = 1	3.6124
K = 2	-1.7822

6.52. The algorithm couldn't run for more than 2 mixtures due to insufficient number of observations in one of the mixtures. Hence the results up to two mixtures are reported for this data. The mixing proportion estimates are given in table 6.53. The mean vectors for each mixture is given in table 6.54. The variances for each group are given by

$$\sigma_1 = 0.2224$$

$$\sigma_2 = 7.7375$$

The classification table is given in table 6.55. The scatter plot matrix for each pair of variables in the Gifi space is shown in figure 6.35. The original data contained 24 patients from the KCS group and 17 patients from the non-KCS group. The algorithm was able to identify two mixtures in the Gifi space where one of the mixture contained 15 patients and the other mixture contained 26 patients. Clearly, one mixture indicates the KCS group and the other mixture indicates the non-KCS group. We do not have any information regarding the prior classification of the observations and hence we are unable to provide the confusion matrix for this data set.

Now, we do a variable selection assuming that the data is generated from two mixtures

Table 6.55: KCSG2 Data: Classification Table

Value	Count	Percent
K = 1	15	36.59%
K = 2	26	63.41%

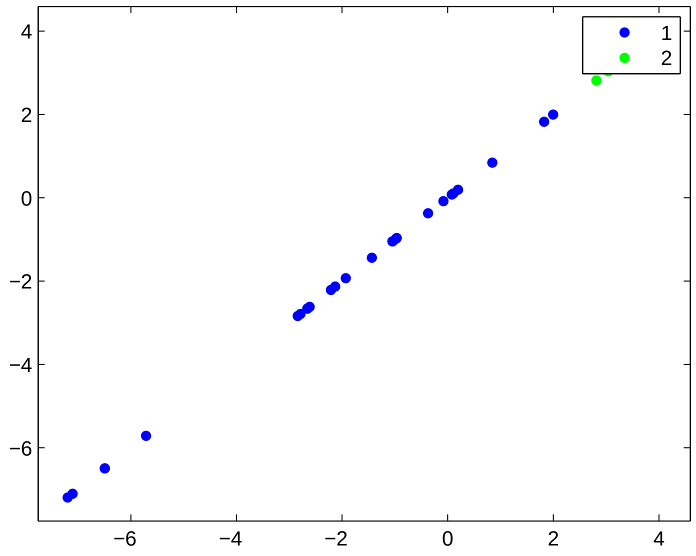


Figure 6.35: KCS Group 2 Data: Scatter Plot Matrix

(since ICOMP is minimum for two mixtures). We use GA for variable selection with maximum iterations of 20, population size of 20, probability of crossover of 0.75, probability of mutation of 0.10 and crossover type as uniform. $ICOMP_{C1}$ is used as the fitness function.

The following parameters are selected by GA process with the above input parameters.

Model : A, B, D, E

Information Criterion Score: -963.4305. The best value of the above fitness function at the end of each iteration of the GA process is shown in the figure 6.36. In the original mixed data space, we could'nt fit more than two mixtures due to insufficient number of observations in one of the mixtures. This data fits one mixture in the original mixed data space and two mixtures in the Gifi space.

Note: In case of insufficient number of observations, one might run into singularity details when computing the inverse of the covariance matrix. In this case, one might use improved covariance smoothers such as maximum entropy, [Theil and Fiebig, 1984], maximum likelihood / empirical bayes, maximum entropy / empirical bayes, stipulated ridge

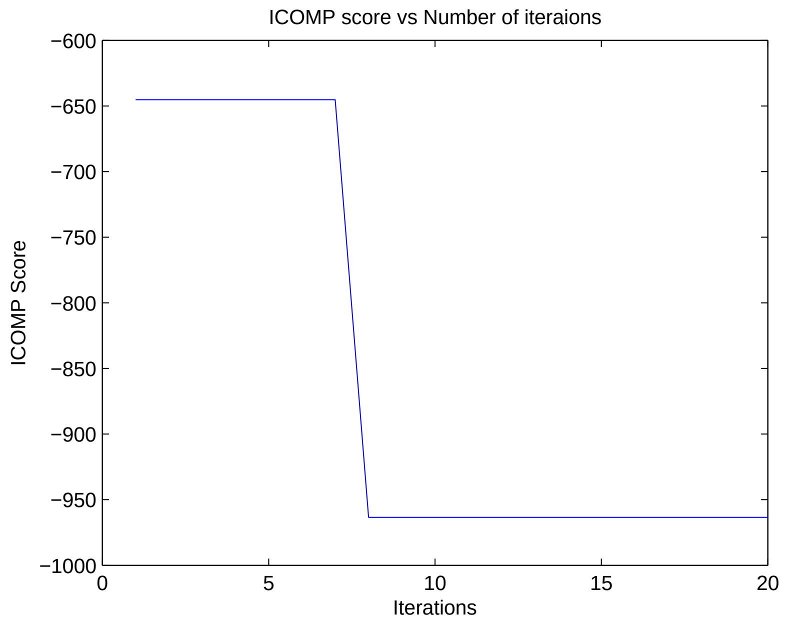


Figure 6.36: KCSG2 (CA): Plot of ICOMP vs Number of iterations in GA

and stipulated diagonal, [Shurygin, 1983], convex sum, [Chen, 1976], shrinkage estimator of ledoit and wolf, [Ledoit and Wolf, 2003].

Chapter 7

Conclusions

In this work, we addressed two problems. The first problem is to determine the optimal number of mixtures in a multivariate Bernoulli distributed data using genetic algorithm and information complexity (ICOMP). We showed that choosing the highest maximum likelihood value by itself does not yield an optimal number of mixtures. We addressed the problem of high dimensional binary data using a genetic algorithm to identify the best set of predictors that are sufficient for classification. We used a slight variation in our GA procedure to the traditional procedure. The results of our GA procedure showed convergence to an optimum solution with minimum number of iterations. We ran our experiments on a simulated data set and also on two real data sets (mobile phone data set and KCS and non-KCS patient data set). The results are explained in detail in the first part of the numerical results section.

The second problem is to mine for some interesting patterns from a mixed data set. We presented the idea of transforming the mixed data space to a continuous space by a mechanism known as Gifi transformation, [Gifi, 1989]. In the Gifi space, the data is purely continuous in nature. Therefore, we can implement the usual multivariate statistical methods on the data in the Gifi space.

We presented two algorithms for implementing the multivariate statistical methods in the Gifi space - the optimal scaling method and the linear combination method. In the optimal scaling method each categorical variable in the Gifi space is optimally scaled by multiplying its indicator matrix with its optimal weight vector. With the optimal scaling method, the data in the Gifi space would be the same as the data in the original space but with the categorical values replaced by its corresponding continuous weight/score values. In the linear combination method, the categorical variables in the Gifi space are collapsed to a 1-dimensional continuous values by the linear combination of the weights of the categories of all the categorical variables in the Gifi space. With the linear combination method, the size of the data in the Gifi space would be different than the data in the original space. The continuous variables would be the same in the Gifi space and the original space. The categorical variables will be transformed to a 1-dimensional continuous values in the Gifi space.

We presented several techniques of the multivariate statistical methods in the Gifi space such as Multiple Regression, Multivariate Regression, Binary Logistic Regression, Multivariate Logistic Regression, Discriminant Analysis and Unsupervised classification. The numerical results showed that the analysis in the Gifi space is very impressive when there are a lot of categorical variables in the data set. We also addressed the problem of high dimensional data using a slight variation of the genetic algorithm with the fitness function, ICOMP.

Bibliography

Bibliography

[Aderberg, 1973] Aderberg, M. R. (1973). *Cluster Analysis for Applications.* Probability and Mathematical Statistics, New York: Academic Press, 1973.

[Akaike, 1973] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. Second international symposium on information theory, Academiai Kiado, Budapest, 267-281.

[Arora and Kannan, 2001] Arora, S. and Kannan, R. (2001). Learning Mixtures of Arbitrary Gaussians. Symposium on Theory of Computing (STOC), 2001.

[Botev, 2005] Botev, Z. I. (2005). A Non-Asymptotic Bandwidth Selection Method for Kernel Density Estimation of Discrete Data. *Department of Mathematics, The University of Queensland, Brisbane, 4072, Australia.*

[Bozdogan, 1987] Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, (No. 52(3)):345–370.

[Bozdogan, 1988] Bozdogan, H. (1988). ICOMP: A new model selection criterion. *In Hans H. Bock (Ed.), Classification and Related Methods of Data Analysis, Amsterdam: North-Holland*, pages 599–608.

[Bozdogan, 1994] Bozdogan, H. (1994). Mixture-Model Cluster Analysis using model selection criteria and a new informational measure of complexity. *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, pages 69–113.

[Bozdogan, 2004] Bozdogan, H. (2004). *Intelligent Statistical Data Mining with Information Complexity and Genetic Algorithms*, chapter 2, pages 15–56. CRC Press LLC.

[Bozdogan, 2005] Bozdogan, H. (2005). Information Complexity and Multivariate Learning Theory: A Computational Approach with Data Mining Applications. *Department of Statistics, The Univeristy of Tennessee, Knoxville, TN, 2005.*

[Bozdogan, 90a] Bozdogan, H. (90a). On the information based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics, Theory and Methods*, (No. 19(1)):221–278.

[Bozdogan, 90b] Bozdogan, H. (90b). Multisample cluster analysis of common principal component model in K groups using an entropic statistical criterion. *Invited paper presented at the International Symposium on Theory and Practice of Classification, December 16-19, Puschino, Soviet Union.*

[Bozdogan and Magnus, 2003] Bozdogan, H. and Magnus, J. R. (2003). Misspecification resistant model selection using information complexity.

[Breiman and Friedman, 1985] Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformation for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, 80:580–619.

[Buuren and Heiser, 1989] Buuren, S. V. and Heiser, W. J. (1989). Clustering n objects into k groups under optimal scaling of variables. *Psychometrika*, 54:699–706.

[Carreira-Perpinan, 2001] Carreira-Perpinan, M. A. (2001). *Continuous latent variable models for dimensionality reduction and sequential data reconstruction.* PhD thesis, Department of Computer Science, University of Sheffield, U.K.

[Carreira-Perpinan and Renals, 2000] Carreira-Perpinan, M. A. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Compuation, 2000.*

[Chen, 1976] Chen, M. (1976). Estimation of covariance matrices under a quadratic loss function. *Research Report S-46*, pages 1–33.

[Dasgupta, 1999] Dasgupta, S. (1999). Learning Mixtures of Gaussians. Proc. of Symposium on Foundations of Computer Science (FOCS), 1999.

[Dasgupta and Schulman, 2000] Dasgupta, S. and Schulman, L. (2000). A Two-Round Variant of EM for Gaussian Mixtures. Conference in Uncertainty in Artificial Intelligence (UAI), 2000.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, 1977.* Series B, 34: 1 - 38.

[der Kooji and Meulman, 1997] der Kooji, A. J. V. and Meulman, J. J. (1997). MURALS: Multiple regression and optimal scaling using alternating least squares. *Softstat '97 Advances in Statistical Software 6, eds. W. Bandilla and F. Faulbaum, Stuggart: Lucius & Lucius*, pages pp. 99–106.

[Ender, 1998] Ender, P. (1998). Hypothesis Testing: Equality of Population Covariance Matrices. UCLA Department of Education.

[Forrest, 1993] Forrest, S. (1993). Genetic algorithms: Principles of natural selection applied to computation. pages 872–878. American Association for the Advancement of Science. Science, 261(2.4).

[Gifi, 1989] Gifi, A. (1989). *Nonlinear Multivariate Analysis.* Wiley Series in Probability and Mathematical Statistics, 1989.

[Goldberg, 1989] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* Published by Addison-Wesley, 1989, New York.

[Gower and Blasius, 2005] Gower, J. C. and Blasius, J. (2005). Multivariate prediction with nonlinear principal components analysis: theory. Quality & Quantity. 39:359–372.

[Groenen et al., 1998] Groenen, P. J. F., Commandeur, J. F., and Meulman, J. J. (1998). Distance analysis of large data sets of categorical variables using object weights. *British Journal of Mathematical and Statistical Psychology*, 51:217–232.

[Guttman, 1941] Guttman, L. (1941). The quantification of a class of attributes: A theory and a method of scale construction. *In P. Horst (Ed.), The prediction of personal adjustment (pp. 319-348).*

[Hastie et al., 1994] Hastie, T. J., Tibshirani, R. J., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89:1255–1270.

[Heiser and Meulman, 1994] Heiser, W. J. and Meulman, J. J. (1994). Homogeneity Analysis: Exploring the distribution of variables and their nonlinear relationships. *In M. Greenacre & J. Blasius (Eds.), Correspondence analysis in the social sciences: Recent developments and applications*, pages pp. 179–209.

[Holland, 1992] Holland, J. H. (1992). Adaptation in Natural and Artificial Systems. pages 66–72.

[Hosmer and Lemeshow, 2000] Hosmer and Lemeshow (2000). *Applied Logistic Regression.* Wiley Series in Probability and Statistics, 2000, second edition.

[Kruskal, 1964] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(115-129).

[Kullback, 1968] Kullback, S. (1968). *Information Theory and Statistics.* Dover Publishers, New York.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. (1951). On information and sufficiency. pages 79–86. Annals of Mathematical Statistcs, 1951.

[Larsen and Liu, 2005] Larsen, D. M. and Liu, J. (2005). Factors affecting clustering of multivariate binary data. Amsterdam, Netherlands. RC33 Sixth International Conference on Social Science Methodology, 2004.

[Ledoit and Wolf, 2003] Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Emperical Finance*, (10(5):603-621).

[Lee, 2007] Lee, S. Y. (2007). *Structural Equation Modeling, A Bayesian Approach.* Wiley Series in Probability and Statistics, 2007.

[Lee et al., 1995] Lee, S. Y., Poon, W. Y., and Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, (No. 48):339–358.

[Lee et al., 90a] Lee, S. Y., Poon, W. Y., and Bentler, P. M. (90a). Full maximum likelihood analysis of structural equation modles with polytomous variables. *Statistics and Probability letters*, (No. 9):91–97.

[Lee et al., 90b] Lee, S. Y., Poon, W. Y., and Bentler, P. M. (90b). A three-stage estimation procedure for structural equation models with polytomous variables. *Pychometrika*, (No. 55):45–51.

[Lindsay, 1995] Lindsay, B. G. (1995). Mixture Models: Theory, Geometry, and Applications. Hayward. Institute of Mathematical Statistics, NSF-CBMS Regional Conference Series in Probability and Statistics Vol. 5.

[Mangano, 1996] Mangano, S. R. (1996). *An Introduction to Genetic Algorithm Implementation, Theory, Application, History and Future Potential.*

[Marczyk, 2004] Marczyk, A. (2004). Genetic Algorithms and Evolutionary Computation. *The TalkOrigins Archive, 2004.*

[McLachlan, 1992] McLachlan, G. (1992). Discriminant Analysis and Statistical Pattern Recognition. *Wiley Series in Probability and Statistics, 1992.*

[McLachlan and Peel, 2000] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley Publishers, 2000.

[Meulman, 1982] Meulman, J. J. (1982). Homogeneity analysis of incomplete data. *Leiden, The Netherlands: DSWO Press.*

[Meulman, 1992] Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57:539–565.

[Meulman, 1993] Meulman, J. J. (1993). Nonlinear principal coordinates analysis: minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46:287–300.

[Meulman, 1996] Meulman, J. J. (1996). Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13:249–266.

[Meulman, 1998] Meulman, J. J. (1998). Optimal scaling methods for multivariate categorical data analysis. *SPSS White Paper*.

[Meulman, 2003] Meulman, J. J. (2003). Prediction and Classification in NonLinear Data Analysis: Something Old, Something New, Something Borrowed, Something Blue. *Physchometrica*, Vol 68(No 4):493–517.

[Meulman and der Kooij, 2000] Meulman, J. J. and der Kooij, A. J. V. (2000). Transformations towards independence through optimal scaling. *Paper presented at the International Conference on Measurement and Multivariate Analysis (ICMMA), Banff, Canada*.

[Meulman et al., 2002] Meulman, J. J., der Kooji, A. J. V., and Babinec, A. (2002). New features of categorical principal component analysis for complicated data sets, including data mining. *In W. Gaul & G. Ritter (Eds.), Classiciation, automation, and new media (pp. 207-217)*.

[Meulman et al., 2004] Meulman, J. J., der Kooji, A. J. V., and Heiser, W. J. (2004). Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. *In D. Kaplan (Ed.), Handbook of qunatitative methodology for the social sciences (pp. 49-70)*.

[Michailidis and de Leeuw, 1996] Michailidis, G. and de Leeuw, J. (1996). The Gifi System of Descriptive Multivariate Analysis. *Statistical Science, 1998*, Vol 13(No. 4):307–336.

[Nierenberg et al., 1989] Nierenberg, D., Stukel, T., Baron, J., Dain, B., and Greenberg, E. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, (No. 130):511–521.

[Nishisato, 1980] Nishisato, S. (1980). Analysis of categorical data: Dual scaling and its applications.

[Nishisato, 1994] Nishisato, S. (1994). Elements of dual scaling: An introduction to practical data analysis.

[Rissanen, 1976] Rissanen, J. (1976). *Minmax entropy estimation of models for vector processes*, chapter System Identification, pages 97–119. Academic Press, New York.

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

[Rissanen, 1989] Rissanen, J. (1989). *Stochastic Complexity in Statistical Inguiriy.* World Scientific Publishing Company, Teaneck, New Jersy.

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

[Shurygin, 1983] Shurygin, A. M. (1983). *The Linear Combination of the Simplestic Discriminator and Fisher's One.* Number pg: 144-158. In Applied Statistics, Moscow, nauka (ed.) edition.

[SPSS, 1999] SPSS (1999). Regression with Optimal Scaling. *SPSS Categories 10.0, eds. J. J. Meulman, W. J. Heiser, and SPSS Inc., Chicago: SPSS Inc.,*, (pp. 1-8, 77-101).

[SPSS, 2004] SPSS (2004). Regression with Optimal Scaling. *SPSS Categories 13.0, eds. J. J. Meulman, W. J. Heiser, and SPSS Inc., Chicago: SPSS Inc.,*, (pp. 1-10, 107-157).

[SPSS, 2006] SPSS (2006). Prediction Accuracy of Regression with Optimal Scaling Transformations: The .632 Bootstrap with CATREG. *Manuscript submitted for publication.*

[Srinivas and Patnaik, 1994] Srinivas, M. and Patnaik, L. M. (1994). Genetic algorithms: a survery. *IEEE Transactions of Signal Processing*, (42(4)):927–935.

[Tamhane and Dunlop, 2003] Tamhane, C. A. and Dunlop, D. D. (2003). *Statistics and Data Analysis from Elementary to Intermediate.* PRENTICE HALL, Upper Saddle River, NJ, USA.

[Theil and Fiebig, 1984] Theil, H. and Fiebig, D. G. (1984). *Exploiting continuity, maximum entropy estimation of continuous distributions.* Number 246 p. Ballinger Publ. Co., Cambridge, Massachusetts.

[Titterington et al., 1985] Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons.

[Trefethen and Bau, 1997] Trefethen, L. N. and Bau, D. (1997). *Numerical Linear Algebra*, pages 56–61. SIAM, Society for Industrial and Applied Mathematics, Philadelphia, USA, 1997.

[VanEmden, 1971] VanEmden, M. H. (1971). *An Analysis of Complexity.* Mathematical Centre Tracts, 1971. Amsterdam.

[Wasserman, 2004] Wasserman, L. (2004). *All of Statistics. A Concise Course in Statistical Inference.* Springler texts in Statistics, 2004.

[Young et al., 1978] Young, F., Takane, Y., and de Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling. *Psychometrika*, 43:279–281.

[Young, 1981] Young, F. W. (1981). Quantitative Analysis of Qualitative Data. *Psychometrika*, 46:357–358.

[Young et al., 1976] Young, F. W., Leeuw, J. D., and Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41:505–529.

Appendix

Appendix A

Appendix

We provide a detail documentation of the Gifi system in this section. The Gifi System was developed using MATLAB 7.4.0 on Windows XP platform running on Pentium (R) 4 CPU 3.00GHz, 504MB of RAM.

A.1 Data Input

The input to the Gifi System must be in MS Excel format. The system automatically detects the header if it finds any non numeric character in the first row of the data file. The initial data input screen of the Gifi System is shown in the figure A.1. The following steps should be followed to import the data from the MS Excel file.

1. Click on the 'Select a file' button. This opens a file open dialog with all the MS Excel files present in the current directory.

2. Select the desired data file and click on the 'Open' button in the file open dialog. The name of the file with the path information is displayed in the text box next to the 'Select a file' button.

3. Click on the 'Import' button. When the data import is complete, the 'Done' button becomes active.

4. Click on the 'Done' button. Now, the user should see the number of variables displayed in the 'Total Number of Variables' text box as shown in figure 2. For example,

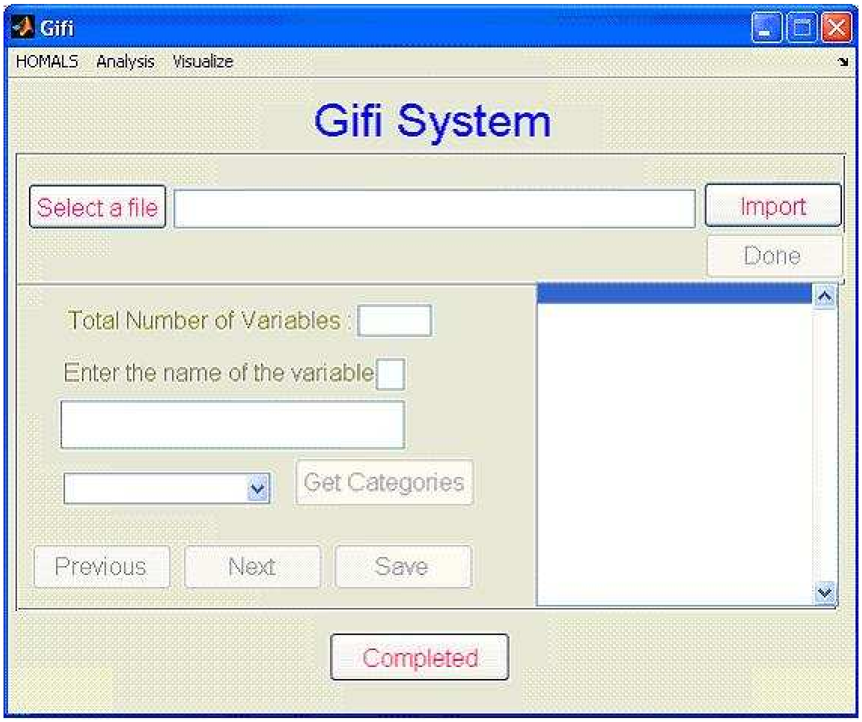


Figure A.1: Gifi System - Initial data input screen

we show the data import of the 'ICU Data'. The variable number and its name are displayed in text boxes as shown in figure A.2.

5. At this point, the user should specify the type of the variable by selecting its type from the combo box shown in figure A.2. If the variable is categorical, the user should select categorical from the combo box. For instance, the variable 'STA' is a categorical variable having two categories '0' and '1'. If we specify a variable as categorical, the 'Get Categories' button becomes active. If the user clicks on the 'Get Categories' button, one can see the categories of that variable in the list box on the right. If the variable is continuous, the user should select continuous from the combo box. For instance, the variable 'AGE' is a continuous variable. If the user selects continuous as the type of the variable, the 'Get Categories' button remains inactive. It becomes active only for the categorical type variables. If the variable is neither categorical nor continuous, the user should leave the combo box unselected. For example, the variable 'ID' is neither categorical nor continuous. Hence we do not select anything in the combo box.

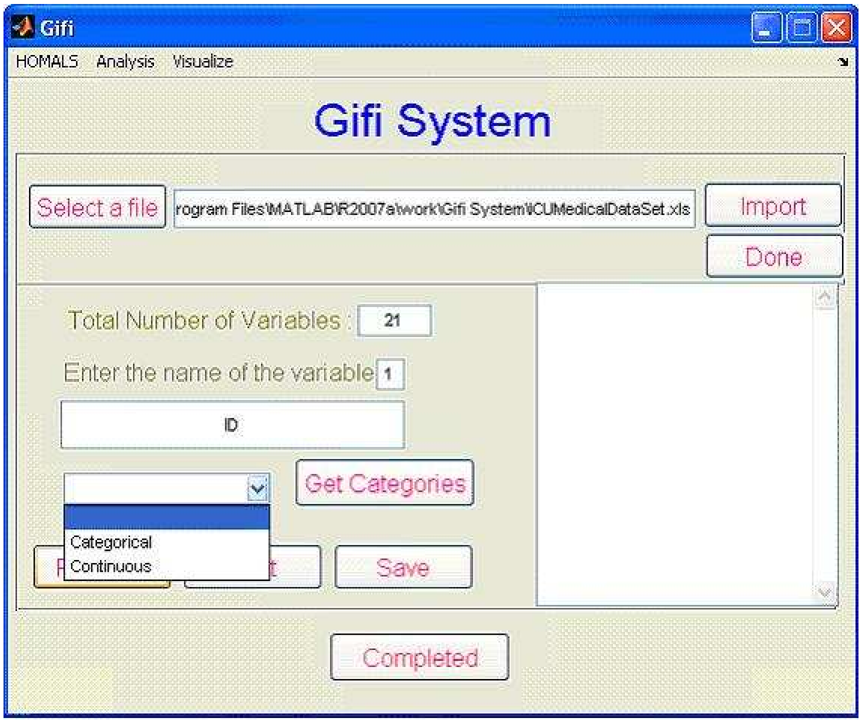

Gifi
HOMALS Analysis Visualize
Gifi System
Select a file
rogram Files\MATLAB\R2007a\work\Gifi System\ICUMedicalDataSet.xls
Import
Done
Total Number of Variables : 21
Enter the name of the variable 1
ID
Get Categories
Categorical
Continuous
Save
Completed

Figure A.2: Gifi System - Intermediate data input screen

6. The user must click on the 'Save' button after specifying the type for each variable.

7. Click on 'Next' button to specify the type of the next variable and repeat steps 5 and 6. In case, if the user has misspecified the type of a previous variable, one can go back by clicking on the 'Previous' button until the specific variable is displayed. At this point, the user can again specify the type of that variable and click on 'Save' button.

8. After specifying the type of all the variables, click on the 'Complete' button. By clicking on this button, all the controls on the panel become inactive. At this point the data and the category specifications have been stored in a structure format in the memory.

A.2 Transformation to the Gifi space

A.2.1 HOMALS - Normalized ALS scores for single solution

1. Go to $HOMALS \rightarrow SingleSolutions \rightarrow Scores$

2. Only the data on the categorical variables are taken for transformation into the Gifi space. The object score quantification (X), the category quantifications (Y), the information loss (sigma), the eigen vector (eta), the discrimination measure (psi), the number of iterations taken for convergence (iter) are reported in the MATLAB editor. Usually the convergence, leads to a unique eigen value but in some cases it might give positive eigen value or a negative eigen value. In this case, we choose the positive eigen value as our arbitrary choice.

3. At this point, the category quantification and the indicator matrices for each categorical variable are stored in their respective structures.

A.2.2 HOMALS - Normalized ALS weights for single solution

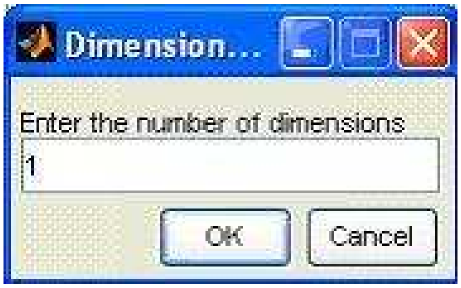
1. Go to $HOMALS \rightarrow SingleSolutions \rightarrow Weights$

2. Only the data on the categorical variables are taken for transformation into the Gifi space. The object score quantification (X), the category quantifications (Y), the information loss (sigma), the eigen vector (eta), the discrimination measure (psi), the number of iterations taken for convergence (iter) are reported in the MATLAB editor. Usually the convergence, leads to a unique eigen value but in some cases it might give positive eigen value or a negative eigen value. In this case, we choose the positive eigen value as our arbitrary choice.

3. At this point, the category quantification and the indicator matrices for each categorical variable are stored in their respective structures.

A.2.3 HOMALS - Normalized ALS scores for multiple solutions

1. Go to $HOMALS \rightarrow MultipleSolutions \rightarrow Scores$ and enter the number of the dimensions in the input dialog box shown in figure A.3 and click on 'Ok' button.



Enter the number of dimensions
1

Figure A.3: Gifi System: Number of dimensions for multiple solutions

2. Only the data on the categorical variables are taken for transformation into the Gifi space. The object score quantification (X), the category quantifications (Y), the information loss (sigma), the eigen vector (eta), the discrimination measure (psi), the number of iterations taken for convergence (iter) are reported in the MATLAB editor. Usually the convergence, leads to a unique eigen value but in some cases it might give positive eigen value or a negative eigen value. In this case, we choose the positive eigen value as our arbitrary choice.

3. At this point, the category quantification and the indicator matrices for each categorical variable are stored in their respective structures.

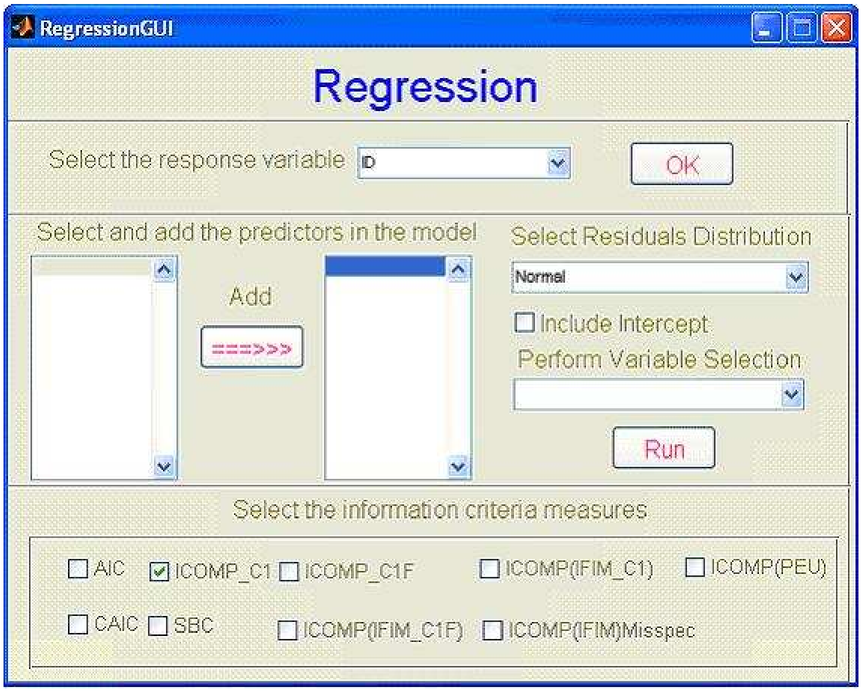
A.3 Analysis

The analysis in the Gifi space can be performed using two methods. They are

- Optimal Scaling Method
- Linear Combination Method

These two methods are described in detail in the 'Applications' chapter of the dissertation.

First, we show the analysis of the Gifi space data using the Optimal Scaling Method (OSM).


RegressionGUI
Regression
OK
Select and add the predictors in the model
Add
===>>>
Perform Variable Selection
Run
Select the information criteria measures

Select the response variable ID

Select Residuals Distribution Normal

☐ Include Intercept

☐ AIC ☑ ICOMP_C1 ☐ ICOMP_C1F ☐ ICOMP(IFIM_C1) ☐ ICOMP(PEU)

☐ CAIC ☐ SBC ☐ ICOMP(IFIM_C1F) ☐ ICOMP(IFIM)Misspec

Figure A.4: Gifi System: OSM Regression screen

A.4 Optimal Scaling Method (OSM)

A.4.1 Regression

1. Select *Analysis* → *Regression*.

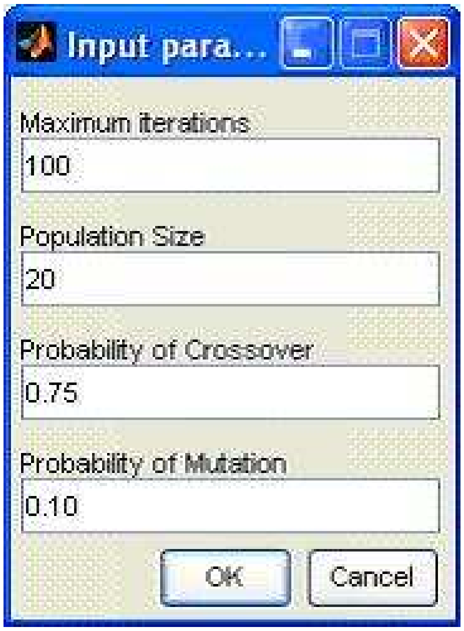
2. The regression GUI will be displayed on the screen as shown in figure A.4.

3. Select the continuous response in the response combo box and click on the ’OK’ button shown in figure A.4.

4. All the variables other than the selected response variable will be displayed in the list box to the left of the ’Add’ button shown in figure A.4.

5. Select the predictor variable in the list box to the left of the ’Add’ button and click on the ’Add’ button. This adds the predictor to the final predictor list and is displayed in the list box to the right of the ’Add’ button.

6. Repeat step 5 until all predictors have been added to the final predictor list.



Maximum iterations
100

Population Size
20

Probability of Crossover
0.75

Probability of Mutation
0.10

Figure A.5: Gifi System: GA input parameters

7. Select the distribution of the residuals from the 'Select Residuals Distribution' combo box shown in figure A.4. By default, the distribution of the residuals is normal.

8. Check the 'Include Intercept' check box.

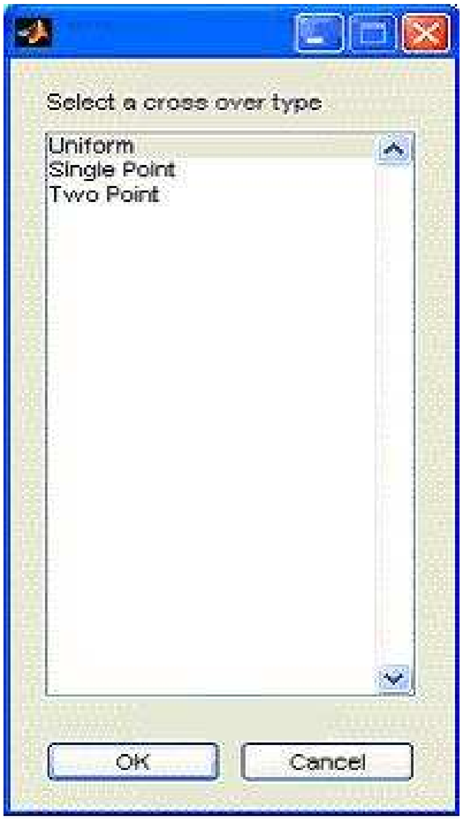
9. Select the information measures from the 'Information Criteria' panel below.

10. Select 'Genetic Algorithm' or 'All Possible Subsets' from the 'Perfrom Variable Selection' combo box. If nothing is selected in the 'Perform Variable Selection' combo box, the system outputs the results for the current model. The current model is one with the response variable and the predictors in the final list of predictor's list box.

11. If the user selects 'Genetic Algorithm' from the 'Perfrom Variable Selection' combo box, the following input dialog shown in figure A.5 appears with default parameters specified.

12. Click on 'OK' button and select the type of crossover from the input list dialog shown below in figure A.6.

13. Select the information criteria from the input list dialog shown in figure A.7.



Select a cross over type

Uniform
Single Point
Two Point

Figure A.6: Gifi System: GA crossover type

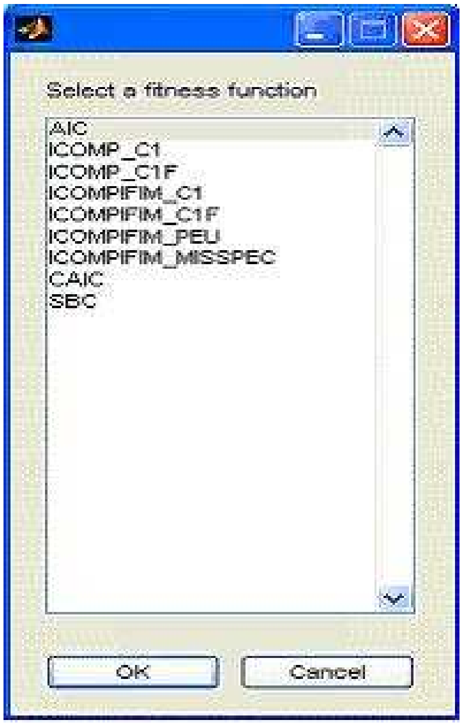


Figure A.7: Gifi System: GA Fitness Function

14. By default the following parameters are selected.

    Maximum Iterations: 100

    Population Size: 20

    Probability of Crossover: 0.75

    Probability of Mutation: 0.10

    Crossover type: Uniform

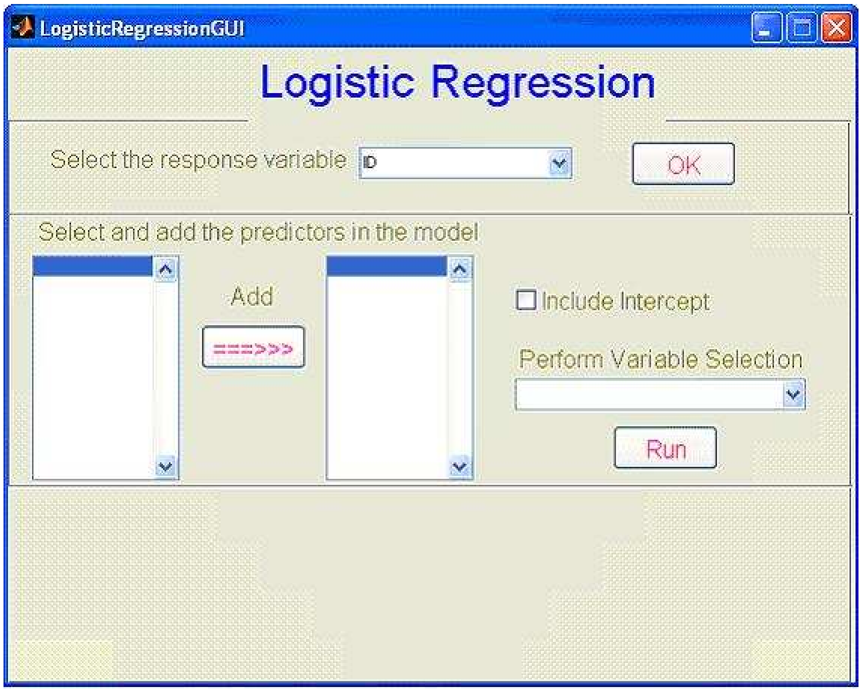
    Information Criteria: AIC

A.4.2 Logistic Regression

1. Go to *Analysis* → *LogisticRegression*.

2. A binary logistic regression GUI is displayed on the screen as shown in figure A.8.

3. Select the binary response variable from the 'Select the response variable' combo box and click on 'Ok' button.



Select the response variable ID

Figure A.8: Gifi System: OSM Logistic Regression Screen

4. All the variables other than the selected binary response variable will be displayed in the list box to the left of the 'Add' button shown in figure A.8.

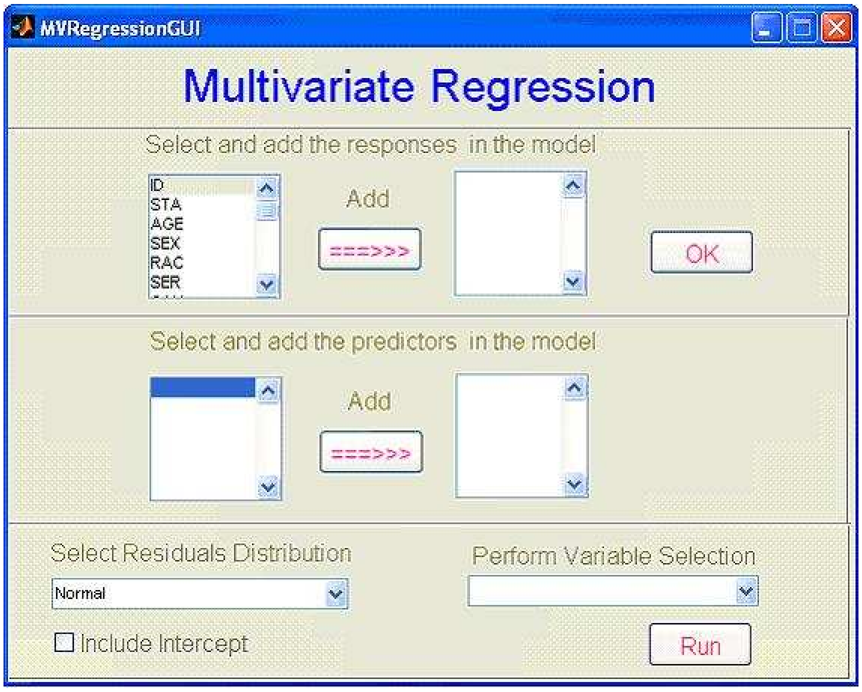
5. Select the predictor variable in the list box to the left of the 'Add' button and click on the 'Add' button. This adds the predictor to the final predictor list and is displayed in the list box to the right of the 'Add' button.

6. Repeat step 5 until all predictors have been added to the final predictor list.

7. Check the 'Include Intercept' check box.

8. Select 'Genetic Algorithm' or 'All Possible Subsets' from the 'Perfrom Variable Selection' combo box. If nothing is selected in the 'Perform Variable Selection' combo box, the system outputs the results for the current model. The current model is the one with the response variable and the predictors in the final list of predictor's list box.

9. If the user selects 'Genetic Algorithm' from the 'Perfrom Variable Selection' combo box, follow steps 11-14 from the Regression procedure under OSM.


MVRegressionGUI
Multivariate Regression
Select and add the responses in the model
ID
STA
AGE
SEX
RAC
SER
Add
===>>>
OK
Select and add the predictors in the model
Add
===>>>
Select Residuals Distribution
Perform Variable Selection
☐ Include Intercept
Run

Select Residuals Distribution: Normal

Figure A.9: Gifi System: OSM Multivariate Regression Screen

A.4.3 Multivariate Regression

1. Go to $Analysis \rightarrow MultivariateRegression$

2. The multivariate regression GUI is displayed on the screen as shown in figure A.9.

3. All the variables in the data set will be displayed in the list box to the left of the response 'Add' button shown in figure A.9. Select each response variable and click on the 'Add' button. This adds the response variable to the final response list displayed in the list box to the right of response 'Add' button.

4. After all the response variables have been added, click on the 'OK' button.

5. All the variables other than the selected response variables will be added to the initial predictor list which is the list box to the left of the predictor 'Add' button shown in figure A.9. Select the predictor variable in the list box to the left of the 'Add' button and click on the 'Add' button. This adds the predictor to the final predictor list and is displayed in the list box to the right of the 'Add' button. Repeat this step until all the predictors have been added.

6. Select the distribution of the residuals from the 'Select Residuals Distribution' combo box shown in figure A.9. By default, the distribution of the residuals is normal.

7. Check the 'Include Intercept' check box.

8. Select 'Genetic Algorithm' or 'All Possible Subsets' from the 'Perfrom Variable Selection' combo box. If nothing is selected in the 'Perform Variable Selection' combo box, the system outputs the results for the current model. The current model is one with the current set of response variables and the predictors in the final list of predictor's list box.

9. If the user selects 'Genetic Algorithm' from the 'Perfrom Variable Selection' combo box, follow steps 11-14 in the Regression procedure under OSM.

A.4.4 Multivariate Logistic Regression

1. Go to $Analysis \rightarrow MultivariateLogisticRegression$.

2. The multivariate logistic regression GUI is displayed on the screen as shown in figure A.10.

3. Follow steps 3-9 in the multivariate regression procedure under OSM.

A.4.5 Discriminant Analysis

1. Go to $Analysis \rightarrow DiscriminantAnalysis$

2. The Discriminant Analysis GUI is displayed on the screen as shown in figure A.11.

3. Select the classification variable from the 'Select the classification variable' combo box and click 'OK' button.

4. Follow steps 4 - 9 from the logistic regression procedure under OSM.

A.4.6 Cluster Analysis

1. Go to $Analysis \rightarrow ClusterAnalysis$.

2. The Cluster Analysis GUI is displayed on the screen as shown in figure A.12.

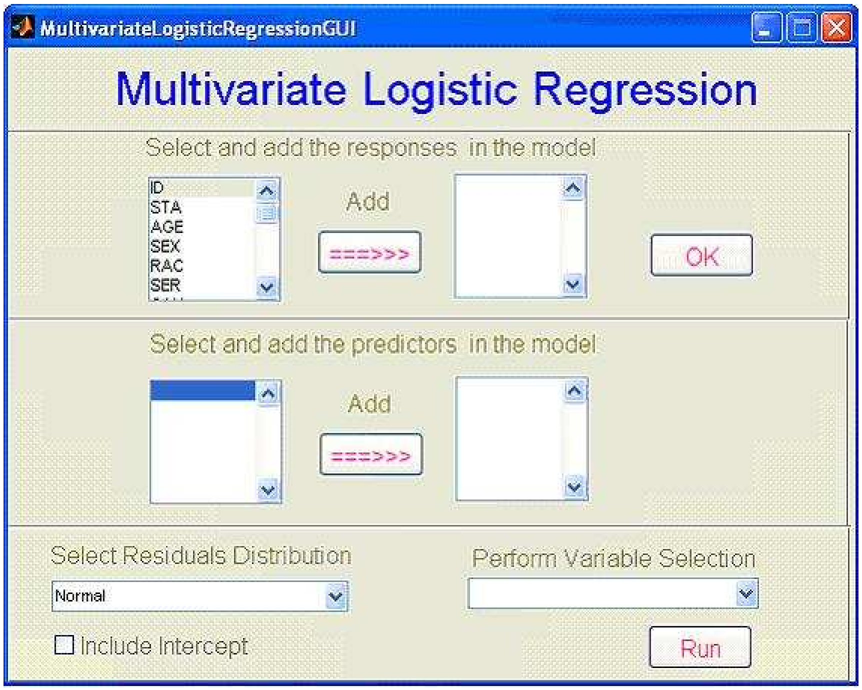

MultivariateLogisticRegressionGUI
Multivariate Logistic Regression
Select and add the responses in the model
ID
STA
AGE
SEX
RAC
SER
Add
===>>>
OK
Select and add the predictors in the model
Add
===>>>
Select Residuals Distribution
Normal
Perform Variable Selection
□ Include Intercept
Run

Figure A.10: Gifi System: OSM Multivariate Logistic Regression Screen

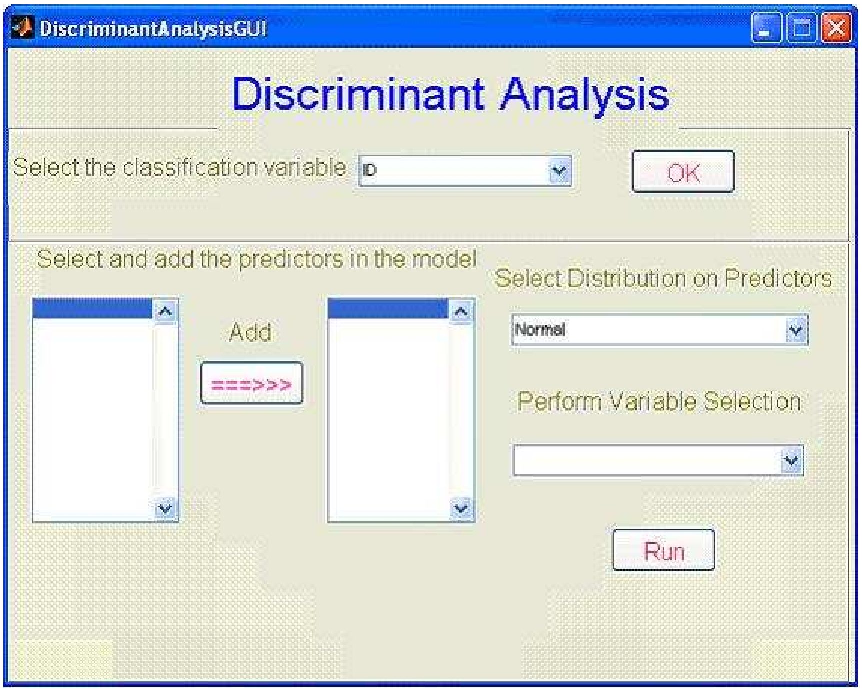

DiscriminantAnalysisGUI
Discriminant Analysis
Select the classification variable
ID
OK
Select and add the predictors in the model
Add
===>>>
Select Distribution on Predictors
Normal
Perform Variable Selection
Run

Figure A.11: Gifi System: OSM Discriminant Analysis Screen

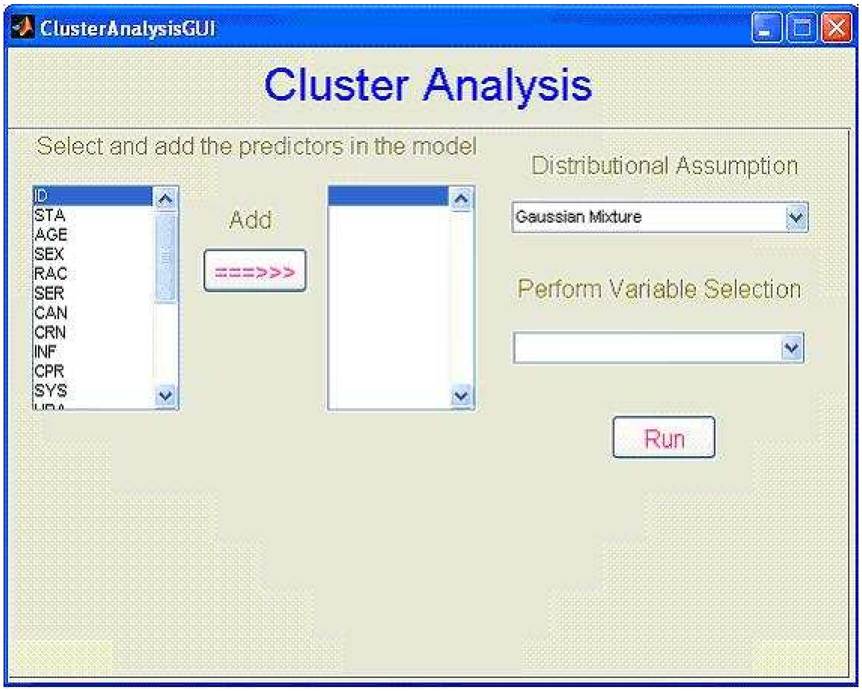


Figure A.12: Gifi System: OSM Cluster Analysis Screen

3. All the variables in the data file are listed in the initial predictor list (list box to the left of the 'Add' button shown in figure A.12.

4. Select the predictor variable in the list box to the left of the 'Add' button and click on the 'Add' button. This adds the predictor to the final predictor list and is displayed in the list box to the right of the 'Add' button. Repeat this step until all the predictors have been added.

5. Select the distribution assumption from the 'Distributional Assumption' combo box.

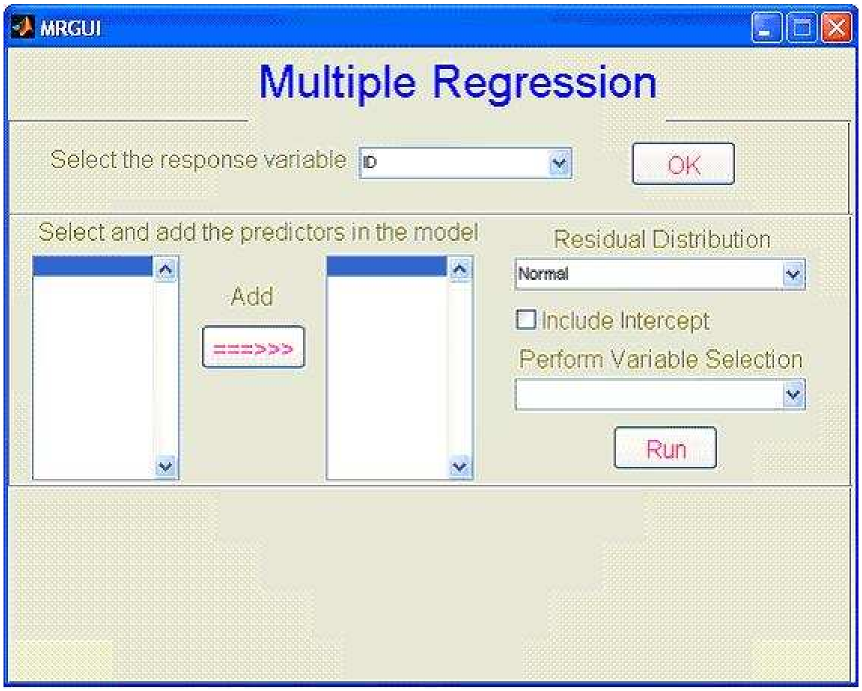
6. Follow steps 8 - 9 in the multivariate regression procedure under OSM.

A.5 Linear combination Method (LCM)

A.5.1 Regression

1. Select $Analysis \rightarrow MR$.

2. The regression GUI will be displayed on the screen as shown in figure A.13.


MRGUI
Multiple Regression
OK
Select and add the predictors in the model
Add
===>>>
Perform Variable Selection
Run

Select the response variable: ID

Residual Distribution: Normal

□ Include Intercept

Figure A.13: Gifi System: LCM Multiple Regression Screen

3. Follow steps 3 - 14 (excluding 9) of the regression procedure under OSM.

A.5.2 Logistic Regression

1. Go to $Analysis \rightarrow BinaryLR$

2. Repeat steps 3 - 9 of the logistic regression procedure under OSM.

A.5.3 Multi-class Logistic Regression

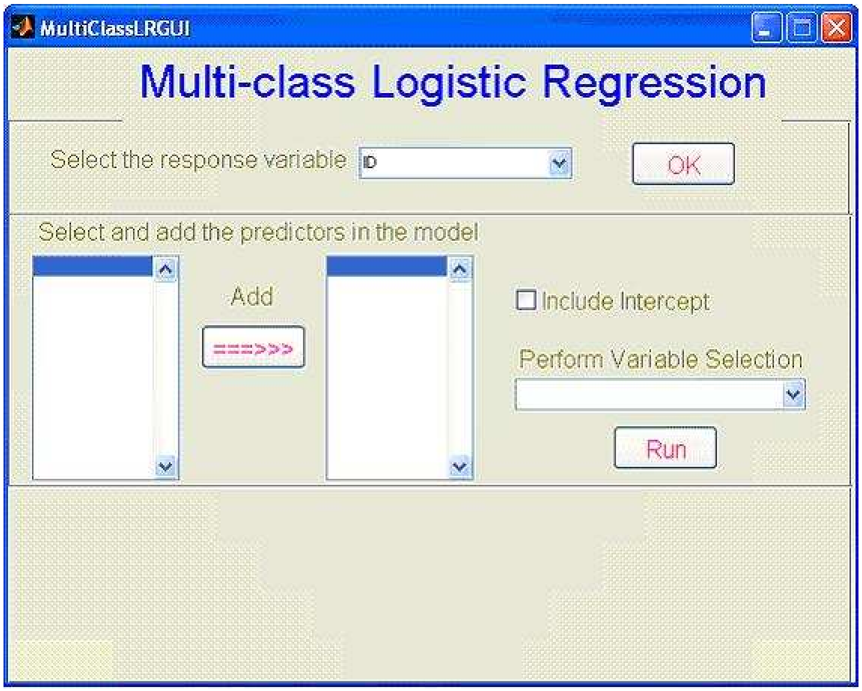
1. Go to $Analysis \rightarrow Multi-classLR$

2. The multi class logistic regression GUI will be displayed on the screen as shown in figure A.14.

3. Repeat steps 3 - 9 of the logistic regression procedure under OSM.

A.5.4 Multivariate Regression

1. Go to $Analysis \rightarrow MVR$


MultiClassLRGUI
Multi-class Logistic Regression
OK
Select and add the predictors in the model
Add
===>>>
□ Include Intercept
Perform Variable Selection
Run

Select the response variable ID

Figure A.14: Gifi System: LCM Multi-class Logistic Regression Screen

2. Repeat steps 3 - 9 in the multivariate regression procedure under OSM.

A.5.5 Multivariate Logistic Regression

1. Go to $Analysis \rightarrow MVLR$

2. Follow step 3 in the multivariate logistic regression procedure under OSM

A.5.6 Discriminant Analysis

1. Go to $Analysis \rightarrow DA$

2. Follow steps 3 and 4 in the Discriminant Analysis under OSM.

A.5.7 Cluster Analysis

1. Go to $Analysis \rightarrow ClusterAnalysis$

2. Follow steps 3 - 6 in the Cluster Analysis procedure under OSM.

Vita

Suman Katragadda was born in Gudiwada, Andhra Pradesh, India in 1982. This place is popularly known as "The Land of Legends" since legendary movie actors like N.T.R and A.N.R, legendary telugu music director Gantasala and the media moghul Ramoji Rao have taken birth here. He was sent to a boarding school, Loyola Public School, at the age of 5 and so stayed away from home for most of the time in his life. Due to his immense interest in quantitative analytics and computation he chose to do a bachelors degree in Computer Science and Information Technology at Vignana Jyothy Institute of Engineering and Technology, JNTU, Hyderabad, AP, India.

In May 2003, Suman matriculated with a B.Tech degree in CSIT and came to the United States in Aug 2003, seeking a masters degree in Computer Science at Kent State University, Kent, OH, USA. At KSU, he happened to meet the world famous researcher in databases, Dr. Yuri Breitbart. With his exceptional performance in the Advanced Database Design course, Dr. Breitbart invited Suman to join him in his research on Cardiological Data Mining project which was a 2 million dollar funded project by the Cleveland Clinic Foundation, Cleveland, OH, USA. During the course of his research he was nominated for the Ohio Board of Reagents award in 2004. After one and a half years of work on this research project he was able to come up with some interesting results that were not known in the medical literature. His results were accepted by the researchers at the Cleveland Clinic Foundation and were published in their journals.

During the course of his masters at Kent State University, Suman was offered a full time job in May 2004 at Axentis Inc, Cleveland, OH, USA. He joined Axentis as an application developer and was later promoted to the rank of software engineer within a few months. He was working at Axentis and on his research project simultaneously from May 2004 to April 2005. He graduated with a master of science degree in Computer Science in May 2005. Due to his immense interest in quantitative field and as per the advice of Dr. Breitbart, he decided to do a masters in statistics. In Aug 2005, he joined The University of Tennessee, Knoxville, TN, USA and it is here where he met another world famous researcher in information statistics, Dr. Hamparsum Bozdogan. He secured an internship from State Farm Insurance in May 2006. It was at State Farm that suman developed his initial fraud detection system to detect suspicious claim estimates, thus saving millions (probably more than a billion according to the Director of Strategic Resources at State Farm) of dollars to State Farm. It was at State Farm where he first noted the problem of modeling mixed data. At the same time, Dr. Bozdogan was also investigating on the same mixed data modeling issue.

Suman decided to do a Ph.D. under the supervision of Dr. Bozdogan on Multivariate Mixed Data Mining. He was awarded the Graduate Excellence Award in Aug 2007 and bagged a place in the students brag book published by the College of Business Administration, UTK, in Summer 2008. Suman intends to complete the requirements for the Ph.D. in Statistics by the end of 2008. His primary research interests are mixed data modeling with applications to evidence based medical data mining in the medical sector, stock market trading in the financial sector and suspicious insurance claims detection in the insurance sector.