5-2009

Data adaptive kernal discriminant analysis using information complexity criterion and genetic algorithm

Dong-Ho Park
University of Tennessee

To the Graduate Council:

I am submitting herewith a dissertation written by Dong-Ho Park entitled "Data adaptive kernal discriminant analysis using information complexity criterion and genetic algorithm." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Education.

Hamparsum Bozdogan and Schuyler W. Huck, Major Professor

We have read this dissertation and recommend its acceptance:

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 Dong-Ho Park entitled "Data adaptive kernel discriminant analysis using information complexity criterion and genetic algorithm." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Education.

Hamparsum Bozdogan, Co-chair

Schuyler W. Huck, Co-chair

We have read this dissertation
and recommend its acceptance:

Jeannine R. Studer

Seong-Hoon Cho

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official students records)

DATA ADAPTIVE KERNEL DISCRIMINANT ANALYSIS USING INFORMATION COMPLEXITY CRITERION AND GENETIC ALGORITHM

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Dong-Ho Park
May 2009

Acknowledgements

This work is dedicated to my son, Jun-sung, who is the most precious one to me. I would also like to dedicate this work to Ji-Young, Yoon-Soo, and Jae-Sung who has been supporting me to finish this work. I would like to thank my advisors, professor Sky Huck and Hamparsum Bozdogan, who have been so supportive of me in my whole doctoral program. Finally, I really appreciate Dr. Andrew Howe's help to proofread my work.

Abstract

This dissertation proposes a new hybrid approach which is computationally effective and easy-to-use for selecting the best subset of predictor variables in discriminant analysis under the assumption that data sets do not follow the normal distribution. Our approach incorporates the information-theoretic measure of complexity (ICOMP) criterion with the genetic algorithm and kernel density estimators in discriminant analysis. This approach enables researchers to find both the optimal bandwidth matrix for the kernel density estimate and the best model from several competing models, which was a severe obstacle for researchers to apply kernel density estimate for discriminant analysis.

The proposed approach is applied to four real data sets and compared with linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and *k*-Nearest Neighbor Discriminant Analysis (*k*-NNDA). Based on our application, we can conclude that our proposed approach performs better than LDA and QDA and performs as well as *k*-NNDA with respect to classification error rates. With our approach we can do all-possible-subset selection of variables for high-dimensional data to determine the best predictors discriminating between the groups.

Keywords and phrases: *Information-theoretic measure of complexity; ICOMP; Kernel density estimate; Variable selection; Subset selection; Model selection; Discriminant analysis; LDA; QDA; NNDA.*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The purpose of this chapter is to introduce an overview of discriminant analysis (DA), research questions and our methodology. The material here is divided into seven sections: (1) Section 1.1 introduces application of DA to education and different methods of DA. (2) Section 1.2 describes model selection in DA and ICOMP. (3) Section 1.3 will give an overview of the genetic algorithm. (4) Then, research questions approached by this dissertation will be provided in Section 1.4. (5) Section 1.5 will explain methods to answer research questions. (6) Section 1.6 will describe the contribution of this dissertation to the DA and model selection. (7) Finally Section 1.7 will explain the organization of this dissertation.

1.1 Overview of Discriminant analysis

Discriminant analysis is popular and widely used in the area of educational research. Some examples of application of DA involve the prediction of the following:

- Academic achievement
- Success in a special education program

• School dropout

• Student success on licensure examination

• Educational placement

There are several different methods in DA. In the next section, we will introduce four different methods.

1.1.1 When data are normal : LDA and QDA

When data conform to the normal distribution, quadratic discriminant analysis (QDA) and linear discriminant analysis (LDA) can be used. Both QDA and LDA are popular and show good performance when data are normally distributed.

LDA assumes sample covariance matrices of each group are all the same. Based on this assumption, it calculates the posterior probability of group membership of each observation, and assigns an observation to a group where the posterior probability of group membership is the greatest. Thus, LDA performs well in homoscedastic cases.

On the other hand, QDA assumes that sample covariance matrices of each group are different. Based on this assumption, it calculates the posterior probability of group membership of each observation, and assigns an observation to a group where the posterior probability of group membership is the greatest. As a result, QDA works well in heteroscedastic cases.

However, LDA and QDA have two major drawbacks, as well. One of the drawbacks is due to small sample size with high-dimensional data. When there are not enough samples, the within-class scatter matrix S_W can be singular. Another problem

occurs when each group does not follow the Gaussian distribution (Qiu & Wu, 2006). If each distribution is not Gaussian, both LDA and QDA are not effective in maximizing the correct classification of group membership, or minimizing the probability of misclassification error rate.

1.1.2 When data are not normal : k-NNDA and KDEDA

In the real world, it is very unlikely that data conform to the normal distribution. Data on one variable may be skewed while data on another variable may have the approximate lognormal distribution, and so forth. As mentioned in the previous section, QDA and LDA are not effective in dealing with data with a nonnormal distribution. There are several approaches to deal with this problem in DA.

One of the popular approaches to handle the problem of nonnormal distributions is k-nearest neighbor discriminant analysis (*k*-NNDA). In *k*-NNDA, the posterior probability of an observation x_i belonging to group *k* is given by:

$$P(k|x_i) = \frac{\pi_k m_k}{\sum_{k=1}^{K} \pi_k m_k}$$

where m_k means the number of observation that are in the neighborhood of the x_i that belong to group k, and π_k is the prior probability of group k.

Qiu & Wu (2006) proposed a new feature extraction method, called a stepwise *k*-NNDA. *k*-NNDA does not depend on the nonsingularity of the within-class scatter matrix, and it does not assume any particular density function. They found that *k*-NNDA

outperforms existing LDA methods, and it was also very efficient, accurate and robust. However, they did not study whether their new method could find an optimal solution.

Another popular approach using nonparametric density estimation is the kernel density estimation approach to discriminant analysis (KDEDA). It uses kernel density instead of normal density assumption in calculating class conditional probability distributions. Lin, Huang and Chang (2004) compared kernel based discriminant analysis with LDA to predict advanced, regular, and remedial placement levels. They found that kernel based discriminant analysis performed better than LDA.

It is widely known that the performance of a kernel density estimator is primarily determined by the choice of a bandwidth, and only in a minor way by the choice of a kernel function (Zhang, King, & Hydman, 2004). In the literature, there is not much work done to choose the optimal bandwidth selection for multivariate kernel (Zhang, King, & Hydman, 2004). This is primarily due to computational difficulty in finding a data adaptive optimal bandwidth matrix.

One approach to find the optimal bandwidth matrix is to use cross-validation methods to minimize misclassification rates for different bandwidth matrices. Sain, Baggerly and Scott (1994) compared the performance of the biased cross validation method, the least-squares cross-validation method, and the bootstrap method for bandwidth selection in multivariate density estimation. They found that the biased cross-validation method performed well compared to the other two methods. However, they also found that the problem of selecting an optimal bandwidth matrix in kernel density

estimation grew in complexity with the dimensionality of data. Cross-validation methods sometimes find multiple values of bandwidth to minimize misclassification rates, from which it is difficult to identify the optimal bandwidth (Ghosh & Bandyopadhyay, 2006).

Zhang, King and Hydman (2004) proposed using Markov chain Monte Carlo (MCMC) algorithms. They treated the elements of the bandwidth matrix as parameters whose posterior density can be obtained through the likelihood cross-validation criterion. They found that the MCMC algorithm generally performed better than the bivariate plug-in algorithm of Duong and Hazelton (2003) and the normal reference rule discussed in Bowman and Azzalini (1997). Yet, they also mentioned that the computation time for higher dimesional data did increase. Increased computational time for high-dimensional data makes its application to discriminant analysis especially impractical.

Bensmail and Bozdogan (2002) compiled eight forms of the bandwidth matrix, and they used Bozdogan's ICOMP (Bozdogan, 1988, 1990, 1994, 2000) as a criterion to choose the optimal bandwith matrix.

This dissertation will focus on data with nonnormal distributions. To handle the problem of nonnormal distributions, the multivariate Gaussian kernel density estimate will be utilized. To choose the optimal bandwidth matrix for the multivariate Gaussia kernel density estimate, eight forms of the bandwidth matrix and ICOMP shown by Bensmail and Bozdogan (2002) will be used.

1.2 Model selection with ICOMP

Model selection and variable selection in discriminant analysis are critical issues. The model selection problem occurs when a researcher needs to choose the best model from several competing potential models. According to Forster (2000), model selection is a bias versus variance trade-off and this is the statistical principle of parsimony. Inference under models with too few variables can be biased, while models with too many variables may provide a poor precision or identification of effects that are, in fact, incorrect. Under the principle of parsimony, researchers prefer a simple model which captures most of the information in data. Moreover, this simple subset model can reduce computational time in subsequent data analysis and reduce undesirable results such as overfitting problem and multicollinearity.

It is well-known that the effect of adding extra variables in multiple regression increases the value of the coefficient of multiple determination, R^2, and cannot decrease it. These redundant variables usually increase model complexity and the positively-biased R^2. In discriminant analysis, according to Huberty & Olejnik (2006), the increase in the number of variables has a different effect compared to multiple regression:

First, unlike regression, it may very well happen that as p increases, the hit rates (separate-group or total-group) will decrease. This is particularly true if the variables to be added do not contribute substantially to the intergroup difference.

Second, similar to regression, as p increases, the positive bias of the internal hit rates (correct classification) increases.

Thus, it is desirable to find the best subset model to develop a rule to increase classification accuracy. The best model without redundant variables may reduce the misclassification error rate and overcome the overfitting problem.

In this dissertation, we will introduce Bozdogan's information-theoretic measure of complexity called ICOMP (Bozdogan, 1988, 1990, 1994, 2000, 2009) as a model selection criterion. ICOMP is based on a generalization of the covariance complexity index originally introduced by Van Emden (1971) and was motivated in part by AIC. ICOMP shows better performance than AIC-type criteria, and it has been applied to multivariate nonnormal regression models (Minhui Liu, 2006), threshould autoregressive models (Kwon, 2003), neural networks and support vector machines (Liu Z. , 2002), and so on. The details of ICOMP will be explained in Chapter 2.

1.3 Genetic algorithm

There are several approaches to find the best subset of independent variables in DA. The all-possible-subset selection method and the stepwise variable selection are common and frequently used methods.

The combinatorial all-possible-subset selection method is effective and guaranteed to find the best model when there are small number of variables. Suppose that there are 5 predictor variables, $X_1, X_2, X_3, X_4,$ and X_5. In this case, we need to analyze 31 $(2^5 - 1)$ models which can be performed without consuming too much time to find the best model. However, it becomes tedious and time-consuming or sometimes impossible to calculate all possible subsets, especially when data is high-dimensional. Suppose that

there are 10 predictor variables. We need to assess 1023 ($2^{10} - 1$) models. The more variables the data has, then the more computational time we need to carry out the analysis. This can make it impossible to use the all-possible-subset selection method to find the best model with high-dimensional data in a reasonable amount of time.

Stepwise variable selection is an alternative approach that can deal with high-dimensional data. There are three types of stepwise variable selection : forward, stepwise and backward. The stepwise variable selection enters variables into equations or removes them from equations based on pre-determined criteria to enter or remove. Widely used statistical packages, such as SAS and SPSS, include stepwise variable selection methods for DA. This may be one of the reasons why the stepwise variable selection method is so popular in DA.

Although the stepwise variable selection method is computationally effective, it has two major problems: (1) the "best subset" model may not emerge, and (2) only one "good" subset of each size is suggested (Huberty & Olejnik, 2006). These are why many seasoned researchers criticize using the stepwise variable selection method in DA and regression analysis as well.

In this research, a Genetic algorithm (GA) will be introduced to choose the best subset of variables. The idea of a GA is based on the Charles Darwin's natural selection in his famous book titled, "On the Origin of Species." According to his theory, individuals that are better adapted survive longer and have a larger probability to mate, thus passing on their variations to the next generation (Schneider & Kirkpatrick, 2006).

Applying this concept to model selection, variables that fit better to equations will be passed on the next generation model. GAs received significant attention through the book of John Holland in 1975, "Adaptation in natural and artificial systems."

A GA is a search technique which is based on principles of natural selection to find optimal or approximate optimal solutions. A GA has two significant advantages: (1) it is independent of the complexity of the problem structure, and (2) it is not likely to be restricted to a local optimal solution (Goldberg, 1989). In addition, the simulation study in Liu (2006) shows that a GA is efficient even when the number of candidate independent variables is large. A GA is used as a variable selection algorithm in regression analysis, and DA (Liu, 2006; Bao, 2004).

1.4 Research Question

Bozdogan's ICOMP has been implemented and has shown superior performance in multiple regression, factor analysis, and classification analysis. However, researchers have not paid much attention to ICOMP for discriminant analysis. Nor have there been studies about incorporating KDEDA, ICOMP, and a GA to handle both nonnormal distributions and high-dimensional data in the area of DA. Therefore, this dissertation has two research questions.

- Whether KDEDA is superior to other methods compared to LDA, QDA, and *k*-NNDA
- Whether the new hybrid approach incorporating KDEDA with a GA using ICOMP is compatible with the all-possible-subset selection approach

1.5 Methods

The purpose of this research is to apply ICOMP as a model selection criterion in KDEDA and to develop an alternative approach in DA to deal with problems of nonnormal distributions and high-dimensional data. ICOMP will be used twice to (1) find the optimal bandwidth matrix in KDEDA (see "1.5.1" below), and (2) find the best model in several competing models (see "1.5.2" below).

1.5.1 The use of ICOMP to find the optimal bandwidth matrix

Selecting an appropriate bandwidth matrix for each model is the most critical factor in the performance of KDEDA (Zhang, King, & Hydman, 2004). To select the optimal bandwidth matrix for a model, eight different bandwidth types tabulated by Bensmail and Bozdogan (2002) will be implemented in KDEDA. The value of ICOMP for each bandwidth type for each group will be calculated for each model. Then, the bandwidth type which minimizes the ICOMP value will be chosen as the optimal bandwidth matrix for the model. For example, suppose there are two variables, X_1, X_2 from three different groups. We calculated ICOMP for all eight bandwidth matrices for each group. These calculated ICOMP values are shown in Table 1.1. Bandwidth type "Σ" is chosen as the optimal bandwidth matrix for each group, because it has the minimum value of ICOMP for each group.

Table 1.1 Example of selecting the optimal bandwidth matrix

Group	λI	$\lambda_k I$	B	λB_k	$\lambda_k B_k$	Σ	Σ_k	NN
1	2894.9	2942.0	2819.7	2949.8	2993.7	2814.8	2974.1	3027.3
2	3606.0	3985.7	3237.1	3653.4	3835.7	3172.8	3768.0	3973.5
3	2820.9	2796.7	2730.2	2763.0	2741.0	2710.2	2725.2	2886.1

1.5.2 The use of ICOMP to find the best model

The next stage is to select the best model among competing models. We will use ICOMP as a model selection criterion for KDEDA. The model with the minimum ICOMP value will be chosen as the best model. For this, we show the derivation of the expression of ICOMP for KDEDA.

While there are various model selection methods, we will implement two approaches, the all-possible-subset selection method and a GA. The all-possible-subset selection method is useful and effective when data set has a small number of variables, while a GA is computationally effective for large data sets with many variables.

For data sets with few variables, we will utilize the all-possible-subset selection method and a GA at the same time. First, the ICOMP value for all possible models will be calculated, and the model which minimizes ICOMP will be chosen as the best model. Second, the GA will be run to find the best model. The GA will find the best model with the minimum ICOMP value. Two results from the all-possible-subset selection method

and the GA will be compared to determine whether the GA is effective in finding the best model.

For data-sets with many variables, a GA will be utilized to choose the best model among competing models. ICOMP will be used as a fitness function in the GA. Then the GA will identify the optimal solution as that with the minimum ICOMP value over different generations. To investigate whether the GA is consistent in finding the optimal solution, we will run the GA 20 times for a data set. If the GA finds one optimal solution in most replications, we may assume that the optimal solution might be the best solution among all possible models. The computational time of KDEDA with the GA will be evaluated to investigate whether the proposed approach is quick enough to be applicable to real world data.

In addition, the performance of KDEDA will be compared with other DA methods such as, LDA, QDA and *k*-NNDA. The classification error rate of a test sample for KDEDA, LDA, QDA and *k*-NNDA will be used as a criterion to evaluate which method is more effective.

1.6 Contribution of this dissertation

This dissertation will make several significant contributions in model selection and DA. First, it will develop the hybrid approach which combines KDEDA, ICOMP, and the GA. We use this hybrid approach to simultaneously choose the optimal bandwidth matrix and the best subset model. Second, the expression of ICOMP for multivariate discriminant analysis will be derived. The new ICOMP expression will be a

significant contribution in the area of model selection. Third, the new approach using KDEDA, ICOMP and a GA will provide a computationally efficient method to find the best discriminating model when data are high-dimensional - without losing the power to find the best or approximate model. Finally, the effectiveness of KDEDA with the GA using ICOMP will be compared with other discriminant analysis methods such as *k*-NNDA, LDA and QDA. This comparison may show the superiority of proposed approach to the other methods.

1.7 Organization of this dissertation

This dissertation consists of six chapters. Chapter one is an introductory chapter. It describes the problems in LDA and QDA, briefly explains the proposed approach, and details contributions to the literature. Chapter two introduces Bozdogan's ICOMP. Chapter three shows a brief explanation of the GA. Chapter four explains KDEDA and model selection in discriminant analysis. This chapter introduces ICOMP as a model selection criterion for bandwidth selection and derives ICOMP as a model selection criterion for the best subset selection of the variables in DA. In chapter five, we apply the new proposed approach to four real data sets. This chapter compares the performance of KDEDA with LDA, QDA and *k*-NNDA. It also shows the performance of the GA in comparison to the all-possible-subset selection method. The final chapter is a summary of the major findings and provides discussion for future research topics.

Chapter 2

ICOMP : Information Complexity Criteria for Model Selection

The purpose of this chapter is to introduce *ICOMP*, the *Information theoretic measure of covariance complexity*, developed by Bozdogan (1988). ICOMP is motivated by AIC and information theory. Therefore, these two concepts are briefly explained to increase the understanding of ICOMP. The majority of this chapter is summarized from Bozdogan's work (Bozdogan, 1988, 1990, 1994, 2000, 2009) on ICOMP.

The material here is divided into five sections: (1) Section 2.1 presents an overview of the AIC and ICOMP. (2) Section 2.2 introduces Shannon's entropy which is one of theoretic foundations of ICOMP. (3) Section 2.3 explains various forms of complexity including $C_0(\Sigma), C_1(\Sigma), C_{\mathrm{F}}(\Sigma)$, and $C_{1\mathrm{F}}(\Sigma)$. (4) Section 2.4 clarifies the definition of ICOMP. (5) Finally Section 2.5 summarizes this chapter.

2.1 The overview of AIC and ICOMP

Akaike's entropy-based information criterion (AIC) introduced in 1973 has had significant impact in the area of model selection. The AIC is based on the concept of

entropy, and it has two components: the lack of fit component, and the penalty component. AIC is given by

$$AIC(k) = \underbrace{-2\log L\left(\hat{\theta}_k\right)}_{lack\ of\ fit} + \underbrace{2k}_{penalty} \tag{2.1}$$

where $L\left(\hat{\theta}_k\right)$ is the maximized likelihood function, $\hat{\theta}_k$ is the maximum likelihood estimate of the parameter θ_k, and k is the number of independent parameters in the model.

The AIC is not a hypothesis testing procedure, and it is different from conventional hypothesis testing procedures. Given a data set, AIC will rank several competing models according to their AIC value. A model with the minimum value of AIC is chosen as the best model to fit the data. Therefore, researchers can get wider inference on the data set based on AIC values of several competing models.

In AIC, the trade-off takes place between a lack of fit term, i.e., $-2\log L\left(\hat{\theta}_k\right)$ and a penalty term, $2k$ which is a measure of complexity. We can also look at this as compensation for the bias in the lack of fit when the maximum likelihood estimators are used (Bozdogan, 2000).

However, several researchers have doubted the validity of penalty term, $2k$. In AIC, estimation bias is corrected by the number of free parameters which is constant and has no variability. Akaike went to asymptotics too quickly when he derived his AIC (Bozdogan, 2000). Hurvich & Tsai (1989) stated that in the case of AIC, there is evidence of an overfitting problem when the dimension of the candidate model increases in comparison to the sample size. When this happens, AIC becomes a strongly negatively-

biased estimate of the information. Rissanen (1976) doubted whether the penalty term, $2k$ is sufficient to prevent overfitting and unnecessary complexity.

The idea of ICOMP was motivated in part by AIC, and in part by information complexity concepts and indices (van Emden, 1971).Compared to AIC, ICOMP is based on the generalization of the information-based covariance complexity index of van Emden (1971). ICOMP is designed to estimate a loss function of a general multivariate linear or nonlinear model.

$$Loss = Lack\ of\ fit + Lack\ of\ parsimony + Profusion\ of\ complexity \qquad (2.2)$$

where profusion of complexity is the measure of dependency or interaction between variables. Estimation of the loss function can be measured by using the additivity property of information theory and the entropic developments of Rissanen (1976) in his final estimation criterion (FEC) in estimation and model identification problems (Bozdogan, 2000). In the next section, I will introduce Shannon's (1951) entropy which is critical to understand the penalty term of ICOMP.

2.2 Shannon's entropy

Covariance complexity is defined here in terms of Kullback-Leibler (1951) information divergence against independence and Shannon's (1951) entropy. Consider the multivariate normal distribution $f(x)$ which is defined by:

$$f(x) = f(x_1, x_2, \dots, x_p)$$

$$= (2\pi)^{-p/2}|\Sigma|^{-1/2} exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \qquad (2.3)$$

where $\mu = (\mu_1, \mu_2, \dots, \mu_p)', -\infty < \mu_j < \infty, j = 1, 2, \dots, p$, $|\Sigma|$ is the determinant of Σ, and $\Sigma > 0$, a positive definite covariance matrix.

According to Blahut (1987, pg 250), the joint entropy $H(x) = H(x_1, x_2, \dots, x_p)$, with arbitrary mean and covariance matrix Σ, is stated by:

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

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

$$= \frac{1}{2}\log(2\pi)^p |\Sigma| + \frac{1}{2} tr\left[\int f(x)\Sigma^{-1}xx'dx\right] \qquad (2.4)$$

Then, since $E[xx'] = \Sigma$,

$$H(x) = H(x_1, x_2, \dots, x_p) = \frac{p}{2}\log(2\pi) + \frac{p}{2} + \frac{1}{2}\log|\Sigma|$$

$$= \frac{p}{2}[\log(2\pi) + 1] + \frac{1}{2}\log|\Sigma| \qquad (2.5)$$

From (2.4), the marginal entropy $H(x_j)$ is given by:

$$H(x_j) = -\int f(x_j) \log f(x_j)\, dx_j$$

$$= \frac{1}{2}\log(2\pi) + \frac{1}{2} + \frac{1}{2}\log(\sigma_j^2) \qquad (2.6)$$

2.3 Information theoretic measure of covariance complexity

In this section, we will introduce various forms of informational complexity of a covariance matrix. This section is summarized from the work of Bozdogan (2007).

2.3.1 Initial definition of covariance complexity: $C_0(\Sigma)$

According to Bozdogan (2000), the complexity of a random vector is a measure of the interaction, or the dependency, between its components. An informational measure of dependence between random variables can be defined in terms of Kullback-Leibler (1951) information divergence against independence and Shannon's (1948) entropy as follows (van Emden, 1971):

$$I(X) = \ I(x_1, x_2, \dots, x_p) = \sum_{j=1}^{p} H(x_j) - \ H(x_1, x_2, \dots, x_p) \qquad (2.7)$$

where $I(X)$ is measure of dependence between random variables, I is the Kullback-Leibler information divergence against independence, $H(x_j)$ is the marginal entropy, and $H(x_1, x_2, \dots, x_p)$ is the joint entropy.

From (2.5) and (2.6), (2.7) can be expressed as follows:

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

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

$$= \frac{1}{2}\sum_{j=1}^{p} \log\left(\sigma_j^2\right) - \frac{1}{2} log|\Sigma| \tag{2.8}$$

where $\sigma_j^2 = \sigma_{jj}$ is the jth diagonal element of Σ and p is the dimension of Σ.

Therefore, the information-theoretic measure of complexity can be defined by:

$$C_0(\Sigma) = I(X) = \frac{1}{2}\sum_{j=1}^{p} \log\left(\sigma_j^2\right) - \frac{1}{2} log|\Sigma| \tag{2.9}$$

$C_0(\Sigma)$ has the following characteristics:

- The first term of $C_0(\Sigma)$ in (2.9) is not invariant under orthonormal transformation
- $C_0(\Sigma) = 0$ if Σ is a diagonal matrix
- $C_0(\Sigma) = \infty$, if $|\Sigma| = 0$

Because of the fact that $C_0(\Sigma)$ is not invariant under orthonormal transformation, $C_0(\Sigma)$ is not effective in measuring complexity between random variables. Equation (2.9) can be improved by using the maximal information theoretic measure of complexity, $C_1(\Sigma)$.

2.3.2 Maximal covariance complexity: $C_1(\Sigma)$

The maximal information theoretic measure of complexity of a covariance matrix Σ from a multivariate distribution is defined by:

$$C_1(\Sigma) = {}^{max}_{T} C_0(\Sigma)$$

$$= \frac{p}{2} log\left[\frac{tr(\Sigma)}{p}\right] - \frac{1}{2}\log|\Sigma| \tag{2.10}$$

where the maximum is taken over an orthogonal transformation T of the overall coordinate system.

$C_1(\Sigma)$ can be expressed in terms of eigenvalues of Σ (Bozdogan, 2000). Suppose $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of Σ, then

$$\frac{tr(\Sigma)}{p} = \bar{\lambda}_a = \frac{1}{p}\sum_{j=1}^{p}\lambda_j \tag{2.11}$$

where $\bar{\lambda}_a$ is the arithmetic mean of the eigenvalues of Σ, and

$$|\Sigma|^{1/p} = \bar{\lambda}_g = \left(\prod_{j=1}^{p}\lambda_j\right)^{\frac{1}{p}} \tag{2.12}$$

is the geometric mean of the eigenvalues of Σ. Then the maximal covariance complexity of Σ can be given by:

$$C_1(\Sigma) = \frac{p}{2}\log\left(\frac{\bar{\lambda}_a}{\bar{\lambda}_g}\right) \tag{2.13}$$

The maximal covariance complexity $C_1(\Sigma)$ has several attractive characteristics:

- $C_1(\Sigma)$ is an upper bound to $C_0(\Sigma)$
- $C_1(\Sigma)$ is the log ratio between the arithmetic mean and the geometric mean of the eigenvalues of Σ
- $C_1(\Sigma)$ is invariant with respect to scalar multiplication and an orthogonal transformation
- $C_1(\Sigma) \to 0$ as $\Sigma \to I_p$. This means that the minimum of $C_1(\Sigma)$ is achieved at the least complex structure
- As interaction between variables increases, the complexity increases. In other words, large values of complexity represent high interaction between the variables, and low values of complexity represents less interaction between the variables

2.3.3Frobenius norm complexity: $C_F(\hat{\Sigma})$, and $C_{1F}(\hat{\Sigma})$

Another measure of complexity of a covariance matrix is based on the Frobenius norm given by (van Emden, 1971):

$$C_F(\hat{\Sigma}) = \frac{1}{s}\|\hat{\Sigma}\|^2 - \left(\frac{tr(\hat{\Sigma})}{s}\right)^2 \tag{2.14}$$

where $\|\hat{\Sigma}\|^2 = tr(\hat{\Sigma}'\hat{\Sigma})$. In terms of the eigenvalues, $C_F(\hat{\Sigma})$ reduces to:

$$C_F(\hat{\Sigma}) = \frac{1}{s}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda}_a)^2 \tag{2.15}$$

$C_F(\hat{\Sigma})$ has the following characteristics:

- $C_F(\hat{\Sigma}) \geq 0$, and $C_F(\hat{\Sigma}) = 0$ when $\lambda_j = \bar{\lambda}_a$
- $C_F(\hat{\Sigma})$ is invariant under an orthogonal transformation. In other words, $C_F(\hat{\Sigma} + kI) = C_F(\hat{\Sigma})$

$C_F(\hat{\Sigma})$ in (2.14) and (2.15) can be expanded by introducing the maximal information complexity, $C_1(\hat{\Sigma})$. Then, the Frobenius norm characterization of the maximal information complexity, $C_{1F}(\hat{\Sigma})$, is given by (Bozdogan, 1988):

$$C_{1F}(\hat{\Sigma}) = \frac{s}{4}\frac{C_F(\hat{\Sigma})}{\left(\frac{tr(\hat{\Sigma})}{s}\right)^2} = \frac{s}{4}\frac{\frac{1}{s}\|\hat{\Sigma}\|^2 - \left(\frac{tr(\hat{\Sigma})}{s}\right)^2}{\left(\frac{tr(\hat{\Sigma}^{-1})}{s}\right)^2}$$

$$= \frac{s}{4}\frac{\frac{1}{s}tr(\hat{\Sigma}'\hat{\Sigma}) - \left(\frac{tr(\hat{\Sigma})}{s}\right)^2}{\left(\frac{tr(\hat{\Sigma}^{-1})}{s}\right)^2} \tag{2.16}$$

In terms of eigenvalues, (2.16) can be given by:

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

$$= \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda}_a)^2 \qquad (2.17)$$

$C_{1F}(\hat{\Sigma})$ has following characteristics:

- $C_{1F}(\hat{\Sigma}) \geq 0$, and $C_{1F}(\hat{\Sigma}) = 0$ when $\lambda_j = \bar{\lambda}_a$
- $C_{1F}(\hat{\Sigma})$ is scale invariant

2.3.4 Example

Consider the famous Fisher's Iris data set, which was introduced by Sir Ronald Aylmer Fisher (1936), as an example of applying discriminant analysis. The data set consists of 150 observations from three different species of Iris flowers. It has four variables which measure sepal and petal lengths and widths. X_1 records the sepal length, X_2 is the sepal width, $X_3$is the petal length, and X_4 is petal width.

If we suppose that$X_1, \ldots, X_4$ are normally distributed with μ and Σ, then the MLE $\hat{\Sigma}$ of Σ is given by:

$$S = \begin{bmatrix} 0.6857 & -0.0424 & 1.2743 & 0.5163 \\ -0.0423 & 0.1900 & -0.3297 & -0.1216 \\ 1.2743 & -0.3297 & 3.1163 & 1.2956 \\ 0.5163 & -0.1216 & 1.2956 & 0.5810 \end{bmatrix}$$

Then, the complexity measure, $C_1(\Sigma)$ and $C_{1F}(\hat{\Sigma})$ can be calculated using (2.10) and (2.17) respectively. $C_F(\Sigma)$ can be obtained by using Matlab function, "norm(S,

'fro')". Table 2.1 shows values of each complexity measure for different subsets of the variables. According to Table 2.1, the values of complexity tend to increase, as the number of variables in the model increase. This is logical, because interactions between variables increase when more variables are included in the model. However, this is not true for all the cases. In some cases, a model with fewer variables than other models may have a larger complexity value. For example, the value of $C_1(\Sigma)$ for the model with two variables

X_1 and X_3 is 0.9762. The value of $C_1(\Sigma)$ for the model with three variables X_1, X_2 and X_4 is 0.8931.

2.4 ICOMP : A new information measure of complexity for model selection

In this section, we will introduce ICOMP as a new model selection criterion to measure the fit between a multivariate structural model and observed data. As mentioned earlier, ICOMP was motivated by in part AIC. AIC penalizes the number of free parameters in the model as shown below:

$$AIC(k) = \underbrace{-2\log L\left(\hat{\theta}_k\right)}_{lack\ of\ fit} + \underbrace{2k}_{penalty} \tag{2.18}$$

However, ICOMP penalizes the covariance complexity of the model. For a multivariate normal linear or nonlinear model, the maximal information-theoretic measure of complexity called ICOMP is defined by:

Table 2.1 Complexity of different models for the iris data

Model	Complexity		
	$C_1(\Sigma)$	$C_{1F}(\hat{\Sigma})$	$C_F(\Sigma)$
X_1	0	0	0.6857
X_1, X_2	0.2001	0.1649	0.7141
X_1, X_3	0.9762	0.4290	3.6646
X_1, X_4	0.5563	0.3356	1.1579
X_2, X_3	0.8662	0.4116	3.1567
X_2, X_4	0.2206	0.1784	0.6450
X_1, X_2, X_3	1.8975	1.1824	3.6995
X_1, X_2, X_4	0.8931	0.7452	1.1875
X_2, X_3, X_4	2.3515	1.2883	3.6999
X_1, X_2, X_3, X_4	3.3973	2.4322	4.2360

$$ICOMP(Overall\ model) =$$

$$\underbrace{-2logL(\hat{\varepsilon}|\hat{\theta})}_{\substack{Lack \\ of\ fit}} + \underbrace{2C_1\left(\widehat{Cov}(\hat{\theta})\right)}_{\substack{Complexity\ due\ to \\ covariance\ matrix\ of \\ the\ parameter\ estimates}} + \underbrace{2C_1(\widehat{Cov}(\hat{\varepsilon})}_{\substack{Complexity\ due\ to \\ covariance\ matrix\ of \\ the\ residual}} \quad (2.19)$$

$$= -2logL(\hat{\varepsilon}|\hat{\theta}) + slog\left[\frac{tr\left(\hat{\Sigma}(\hat{\theta})\right)}{s}\right] - log\left|\hat{\Sigma}(\hat{\theta})\right| + nlog\left[\frac{tr\left(\hat{\Sigma}(\hat{\varepsilon})\right)}{n}\right] - log\left|\hat{\Sigma}(\hat{\varepsilon})\right| \quad (2.20)$$

The first part of ICOMP in (2.19),$-2logL(\hat{\varepsilon}|\hat{\theta})$, measures the lack of fit of the model. The second term in ICOMP, $2C_1\left(\widehat{Cov}(\hat{\theta})\right)$, measures the maximal complexity of covariance matrix of the parameter estimates. The third part of ICOMP, $2C_1(\widehat{Cov}(\hat{\varepsilon})$, measures the maximal complexity of the covariance matrix estimated from the model residuals. ICOMP will choose a model with the minimum score as the best model among competing models. For proof of (2.20), we refer the readers to Bozdogan (2000).

Another approach to ICOMP is to use the estimated inverse-Fisher information matrix (IFIM). This approach derives ICOMP as an approximation to the sum of two Kullback-Leibler (KL) distances. For a multivariate model, the general form of ICOMP(IFIM) is defined by:

$$ICOMP(IFIM) =$$

$$\underbrace{-2\log L\left(\hat{\theta}\right)}_{Lack\ of\ fit} + \underbrace{2C_1\left(\hat{F}^{-1}\right)}_{\substack{Complexity\ of\ the\ accuaracy\ of\ the \\ parameter\ estimates\ of\ the\ model\ as \\ measured\ by\ IFIM\ (inverse\ Fisher\ information\ matrix)}} \quad (2.21)$$

$$= -2\log L\left(\hat{\theta}\right) + 2\,\frac{s}{2}\log\left[\frac{tr\left(\hat{F}^{-1}\right)}{s}\right] - \frac{1}{2}\log\left|\hat{F}^{-1}\right| \quad (2.22)$$

where $L(\hat{\theta})$ is the maximized likelihood function, and $\hat{\theta}$ is the maximum likelihood estimate of the parameter θ, $C_1(\hat{F}^{-1})$ is the maximal information complexity of the $\hat{F}^{-1}$, estimated IFIM and $s = rank(F^{-1})$. The first part of ICOMP in (2.21) measures the lack of fit of the model, and the second part of ICOMP in (2.21) measures the maximal complexity of the estimated IFIM. For proof of (2.21), we refer readers to Bozdogan (2000).

2.5 Conclusion

In this chapter, we briefly introduced AIC and ICOMP. Both AIC and ICOMP have had a significant impact on the theory and practice of model selection, and they have their own unique characteristics.

AIC provided an innovative idea in the model selection area. One of advantages of AIC is that it is easy to apply, because it penalizes the model complexity in terms of the number of free parameters. However, as several researchers have mentioned, AIC often overfits the model - especially when the dimension of the candidate model is large in comparison to the sample size.

Compared to other AIC-type criteria, ICOMP has several different characteristics. (1) ICOMP measures the fit between multivariate structural models and observed data as an example of the application of the covariance complexity measure. (2) ICOMP measures dependency between the random variables in the model. (3) ICOMP penalizes the covariance complexity instead of the number of free parameters in the model.

Therefore, the penalty term in ICOMP is more robust than that of AIC, or AIC-type criteria. In addition, numerical examples in model selection, prediction and perturbation studies (Bozdogan, 2000) clearly demonstrate the excellent performance of ICOMP class criteria compared to AIC.

Chapter 3

Genetic Algorithm

In this chapter, the Genetic Algorithm (GA) for model selection will be introduced. As mentioned earlier, the idea of the GA is based on natural selection as described in Charles Darwin's famous book, "On the Origin of Species." According to this theory, individuals that are better adapted survive longer and have a larger probability to mate, thus passing on their variations to the next generation (Schneider & Kirkpatrick, 2006). Applying this concept to model selection, variables that fit better to equations will be passed on the next generation. GAs received significant attention, in part due to the 1975 book of John Holland, "Adaptation in natural and artificial systems." Currently, GAs are widely used in the area of financial management, manufacturing scheduling, chemistry, astronomy, and other areas of data mining. For more information, readers are referred to Goldberg (1989) or Michalewicz (1992). Goldberg's GA is summarized in the following sections.

3.1 An overview of a GA

A GA starts with a population of *solutions* called a *generation*. A solution is represented by a binary string called a *chromosome*. Each solution represents a potential model and can be thought of as an *individual* in the population. For example, for

subsetting in discriminant analysis with $p = 5$, we need to encode each solution as strings with 5 binary codes. A solution can be encoded as a "1 0 0 0 1" which represents a model including variables 1 and 5, and excluding variables 2, 3, and 4.

There are two important aspects to which we need to pay attention in creating an initial population. First, it is typically created randomly to eliminate selection bias. Solutions in the initial population have a significant effect on finding the best solution. Second, the population size N is an important parameter of a GA (Bozdogan, 2004). Population size N determines the number of chromosomes in a population. If there are too few chromosomes, a GA has a few possibilities to perform crossover, though the computational time is fast. This will reduce the possibility of finding the approximate optimal solution. On the other hand, if there are too many chromosomes, a GA slows down, and there is a high possibility that it will find the approximate optimal solution.

A GA uses a criterion called a *fitness function* to evaluate each chromosome in each population. There are many model selection criteria available such as AIC, BIC, CAIC, and so on, that could fill this role. In this research, Bozdogan's ICOMP will be used as a fitness function. In next chapter, we derive the expression of ICOMP for multivariate discriminant analysis.

After scoring, the next step is to select pairs of solutions in the current population to breed a new generation. Although there are many available selection methods, the roulette wheel selection approach is used in this research. The roulette wheel selection approach is popular and is analogous to natural selection. Accodring to the roulette wheel

selection method, the probability of a solution being selected is proportional to its fitness value. Suppose the sample population of 4 chromosomes in Table 3.1 Table 3.1 shows the fitness value and the slection probability of each chromosome. The selection probability is calculated by dividing each fitness value by the total fitness value of population. For example, chromosome 1 has 10% of probability of being chosen, and chromosome 4 has 40% of probability of chosen. Chromosome 4 has 4 times higher probability of being chosen. Therefore, the fitter solutions have higher probability of being selected. This approach is called the roulette wheel selection approach, because the selection probability is analogous to the probability of winning a roulette wheel game.

Elitism is anothery type of selection method. Elitism guarantees that the best individual in a current population is transferred to the next generation. The rest of the individuals for the next generation will be chosen based on the other selection methods such as the roulette method. Elitism has a significant effect on the performance of a GA,

Table 3.1 Fitness value and selection probability

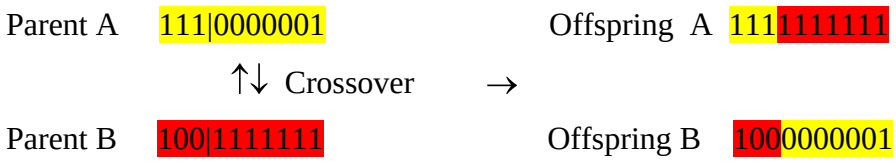
No	Chromosome	Fitness value	Selection Probability
1	10000	100	.1
2	01000	200	.2
3	01100	300	.3
4	01110	400	.4
Total		1000	1

because it always carry the best individual in current population to the next generation of population.

After a pair of individuals are selected, they are used to generate a pair in the next generation called *offsprings*. This reproductive process is performed by GA operators: *crossover* and *mutation*. Crossover is similar to the biological mating process, and the varying portion of chromosomes of parents is controlled by a *crossover probability*. Having a crossover probability of zero means that there is no crossover between chromosomes in the mating pool, and the offsprings are exact copies of their parents. Conversely, a crossover probability of one means that crossover between all chromosomes in the mating pool will always occur.

There are several different types of crossover operations. The most common three types of crossover will be introduced here. In what follows, '|' represents a crossover point where the chromosomes are broken into two portions for crossover.

- **Single point crossover**: a single crossover point is picked. The information beyond this point in either chromosome is interchanged. The resulting chromosomes are the offsprings.



- **Two point crossover :** two crossover points are selected randomly. All the bits between two crossover points are switched between two parents.

Parent A 111|0000|001 Offspring A 1111111001

↑↓ Crossover →

Parent B 100|1111|111 Offspring B 1000000111

- **Uniform crossover :** bits in each chromosome are randomly switched between parents with a fixed probability.

Parent A 1110000001 Offspring A 1100000011

↑↓ Crossover →

Parent B 1001111111 Offspring B 1011111101

Mutation is another type of genetic operator that changes a certain arbitrary bit in a chromosome. Mutation is controlled by a mutation rate or a mutation probability. For example, a '1' can be changed to a '0' by mutation, meaning that a certain predictor variable is either included or excluded from the model. Mutation has a significant effect on the GA by changing the value of bits in the chromosomes. This has the effect of increasing diversity of the considered models, so that the GA can expand the search beyond a locally optimal solution.

Researchers can change the degree of chromosomal modifications by changing the probability of crossover and mutation. The next generation will be more different than the current generation of population, when the crossover rate and the mutation rate are higher.

The next generation is then evaluated based on the fitness function, and will be used to generate the third generation.

The GA continues to produce a next generation until it satisfies certain termination conditions. Common termination conditions include fitness threshold and generation numbers. A GA stops when the number of generations exceeds a pre-specified number of generations, or when the evaluated value of a fitness function exceeds preset value.

In general, the general procedure of the GA can be summarized in six steps as follows:

1. Initialization: Randomly generate a population of N solutions. Populations are chosen by random rule to eliminate selection bias in generating the initial population.
2. Fitness: Evaluate each individual (or model) in the population, based on a model selection criterion which is called a fitness function.
3. Selection: Based on the fitness value, select two individuals from the current population as parents to breed a pair of offspring.
4. Reproduction: Generate new offspring from parents, using two genetic operators: crossover and mutation
5. Elitism: if required.
6. Replace: Place generated offspring in a new population.
7. Repeat steps 2 through 6 until a certain termination condition is satisfied.

3.2 Application of ICOMP in the GA

In this dissertation, we use the ICOMP as the fitness function in the GA for the KDEDA. This approach was proposed by Bozdogan (2004) for the regression model, and it can be applied to KDEDA in the same way. The overall procedure to select the fitter models for mating can be summarized as follows.

First, calculate the ICOMP value for each of the possible subset models in the population

Second, subtract the ICOMP value of each model from the maximum ICOMP value in the population.

$$\Delta\, ICOMP_{(i)} = ICOMP_{(Max)} - ICOMP_{(i)} \tag{3.1}$$

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

Third, average these differences.

$$\overline{\Delta ICOMP} = \frac{1}{N}\sum_{i=1}^{N} \Delta ICOMP_{(i)} \tag{3.2}$$

Fourth, calculate the ratio of each model's difference value to the mean difference value.

$$\frac{\Delta ICOMP_{(i)}}{\overline{\Delta ICOMP}} \tag{3.3}$$

The ratio in equation (3.3) is used to select models which will be included in the mating pool. When the ratio of a model is higher than that of other models, it has a higher chance of being selected. For example, a model with a ratio of two is twice as likely to be selected as a model with a ratio of one. This selection process continues until the number

of offspring equals the initial population size. Table 3.2 illustrates how to calculate the selection probability of a model.

3.3 Advantages and disadvantages of the GA

The GA has several significant advantages compared to other conventional optimization methods.

First, the GA is simple and easy to implement. It only needs a fitness function and does not require additional auxiliary information, such as gradients. Therefore, the GA is useful to solve complex problems or ill-conditioned problems.

Second, the GA is a global optimization search method. Finding the global optimum is more challenging than finding local optima. However, due to crossover and

Table 3.2 ICOMP as a fitness function and the selection probability

No	Chromosome	*ICOMP*	$\Delta ICOMP$	$\Delta ICOMP_{(i)} / \overline{\Delta ICOMP}$	Selection Probability
1	10000	100	300	2	0.50
2	01000	200	200	1.3	0.33
3	01100	300	100	0.67	0.17
4	01110	400	0	0	0
Total		1000	600	3.97	1

mutation process, the GA can overcome local optima, and find the global optimum. Crossover allows for the exchange of information between different models. Changed information may increase the fitness value of a model, thus enabling a GA to move from local optima to the global optimum. Mutation has a more significant effect than crossover on overcoming local optima. If an entire population has converged to a local optimum, crossover can do little to maximize the fitness of parameter - the information exchanged is almost identical. Clearly, crossover of almost identical chromosomes will produce almost identical offspring with nearly identical fitness values. However, mutation allows the GA to produce offspring with genetic segments that are totally different from that of the parents. These new offspring may be in the vicinity of the global optimum (Williams, 2005).

Third, the GA can reduce computational time. Though the GA is not mathematically guaranteed to find the global optimum, it usually finds acceptably good solutions to problems without calculating all the possible models. Reducing computational time can be a critical factor when data sets have many variables and observations.

Of course, the GA also has disadvantages. As already stated, it is not mathematically guaranteed to find a global optimal solution. The GA sometimes converges on sub-optimal solutions. This is likely to occur when several highly fit individuals dominate the population, and the influence of the optimal individual is trivial.

In this case, these several highly fit individuals force the GA to remain in the sub-optimal solution.

Chapter 4

KDEDA and Model Selection

The purpose of this chapter is to introduce KDEDA and propose a new hybrid approach which will combine the GA, KDEDA and ICOMP. The material here is divided into 5 sections: (1) Section 4.1 introduces discriminant analysis. (2) Section 4.2 briefly explains linear and quadratic discriminant analysis. (3) Section 4.3 addresses the *k*-nearest neighbor discriminant analysis (*k*-NNDA), which is a nonparametric discriminant analysis method. (4) Section 4.4 explains the KDEDA and shows several examples about how to choose the optimal bandwidth matrix. (5) Finally, Section 4.5 derives the expression of ICOMP for KDEDA, and proposes the new hybrid approach which combines the GA, KDE and ICOMP for discriminant analysis. The pseudo code for the proposed approach is provided.

4.1 Overview of Discriminant Analysis

DA is one of the popular multivariate statistical methods. In DA, we want to classify an observation into mutually exclusive groups (or classes) based on a certain rule which is called the discriminant function. The goal of discriminant analysis is to minimize the error rate of misclassification when we predict group membership of each observation. The process of discriminant analysis is similar to that in multiple regression

analysis, where one is typically predicting a score on a continuous variable instead of predicting group membership. In both situations, a rule based on a given data matrix is developed and may be used with new observation (from test sample) to predict group membership or scores (Huberty & Olejnik, 2006).

DA is popular and widely used in the area of educational research. Some examples of the application of DA involve predicting the following:

- Academic achievement
- Success in a special education program
- School dropout
- Student success on licensure examination
- Gifted education and talent development
- Educational placement

There are several different methods in DA. In the next section, we will explain four different methods.

4.2 Linear and Quadratic discriminant analysis

Consider that we have samples from k populations, and each group has a size of n_k, $k = 1,2,\ldots, K$ on p variables. A data matrix mentioned above can be given by:

$$X = \begin{bmatrix} x_{111} & x_{121} & \cdots & x_{1p1} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{11n_1} & x_{12n_1} & \cdots & x_{1pn_1} \\ x_{211} & x_{221} & \cdots & x_{2p1} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{21n_2} & x_{22n_2} & \cdots & x_{2pn_2} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{K11} & x_{K21} & \cdots & x_{Kp1} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ x_{K1n_K} & x_{K2n_K} & \cdots & x_{Kpn_K} \end{bmatrix} \quad (4.1)$$

To classify an observation into a group or class, the Bayes' rule is utilized. Let $f_k(\mathbf{x})$ represent the conditional density of an observation vector **x,** when **x** comes from group k. The posterior probability for observation x_i, which belongs to group *k*, is stated by the Bayes' rule:

$$P(K = k \,|\mathrm{x}_i) = \frac{f_k(\mathrm{x}_i)\pi_k}{f(\mathrm{x}_i)} = \frac{f_k(\mathrm{x}_i)\pi_k}{\sum_{k=1}^{K} \pi_k f_k(\mathrm{x}_i)} \quad (4.2)$$

An observation is classified into the group or class such that the posterior probability of group membership is the greatest.

In calculating the posterior probability, we assume that the class conditional density is the multivariate normal distribution, given by

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\Sigma_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)} \quad (4.3)$$

where Σ_k is the $p \times p$ population covariance matrix, $|\Sigma_k|$ is the determinant of Σ_k, which is called the generalized variance of the set of p variables, and $\boldsymbol{\mu}_k$ is the $p \times 1$ population mean vector.

In practice, the parameters Σ_k, and $\boldsymbol{\mu}_k$ in (4.3) are not known, so we use the maximum likelihood estimators (MLE). The MLE of Σ_k is given by:

$$\mathbf{S}_k = \hat{\Sigma}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbf{x_{ki}} - \bar{\mathbf{x}}_k)(\mathbf{x_{ki}} - \bar{\mathbf{x}}_k)' \quad (4.4)$$

where $\mathbf{S}_k$ is the $p \times p$ sample covariance matrix for group k. The MLE of $\boldsymbol{\mu}_k$ is given by:

$$\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbf{x_{ki}}) \quad (4.5)$$

where $\bar{\mathbf{x}}_k$ is the $p \times 1$ vector of sample mean for group k.

After we insert these estimates into expression (4.3), the multivariate normal distribution can be written as:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{S}_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}_k})'\mathbf{S}_k^{-1}(\mathbf{x}-\bar{\mathbf{x}_k})} \quad (4.6)$$

In terms of classification rule, an observation is classified into a group where the posterior probability of group membership is the greatest. Therefore the denominator in (4.2) can be ignored because the value of the denominator is equal for all groups, and does not have any effect on the order of posterior probabilities for each group. The classification rule can be written as:

$$P(K = k\,|\mathbf{x}_i) = f_k(\mathbf{x}_i)\pi_k \quad (4.7)$$

Equation (4.7) can be expressed in terms of natural logarithm as follows:

$$log\,P(K = k \mid \mathrm{x}_i) = \; log f_k(\mathrm{x}_i) \; + \; log\pi_k \tag{4.8}$$

Substituting from equation (4.6), the log posterior probability can be given as:

$$log\,P(K = k \mid \mathbf{x}_i) = \; -\tfrac{1}{2}\left(\mathbf{x} - \bar{\mathbf{x}}_k\right)' \mathrm{S}_k^{-1}\left(\mathbf{x} - \bar{\mathbf{x}}_k\right) \; - \; \tfrac{1}{2} log|\mathbf{S}_\mathrm{k}| - \tfrac{p}{2} log(2\pi) \; + \; log\,(\pi_k) \tag{4.9}$$

In equation (4.9), $-\frac{p}{2} log\,(2\pi)$ is equal for all groups. Therefore, it can be ignored for the classification purpose. The above equation can be written as:

$$log\,P(K = k \mid \mathbf{x}_i) = \; -\tfrac{1}{2}\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right)' S_k^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right) \; - \; \tfrac{1}{2} log|\mathbf{S}_\mathrm{k}| + log(\pi_k) \tag{4.10}$$

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

$$Q_k(\mathbf{x}_i) > Q_l(\mathbf{x}_i) \tag{4.11}$$

for all $k \neq l$, where

$$Q_k(\mathbf{x}_i) = -\tfrac{1}{2}\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right)' S_k^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right) \; - \; \tfrac{1}{2} log|\mathbf{S}_\mathrm{k}| + log(\pi_k) \tag{4.12}$$

Classifying observations based on the values of $Q_k(\mathbf{x}_i)$ is called quadratic discriminant analysis (QDA).

Linear discriminant analysis (LDA) is a special case of QDA, in which we consider that the population covariance matrices for each group are equal. In this case, the sample covariance matrices are equal for all groups: $\mathbf{S} = \mathbf{S}_\mathrm{k}$. Then, the equation (4.12) above can be written as:

$$L_k(\mathbf{x}_i) = -\tfrac{1}{2}\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right)' \boldsymbol{S}^{-1}\left(\mathbf{x}_i - \bar{\mathbf{x}}_k\right) \; - \; \tfrac{1}{2} log|\mathbf{S}| + \log(\pi_k) \tag{4.13}$$

After some matrix manipulation, we can find that $-\frac{1}{2} log|\mathbf{S}|$, and $-\frac{1}{2}\mathrm{x}_i{}'\boldsymbol{S}^{-1}\mathrm{x}_i$ are

common for all groups, and they may be ignored for classification purposes. Then equation (4.13) can be written as:

$$L_k(\mathbf{x}_i) = \bar{\mathbf{x}}_k' \boldsymbol{S}^{-1} \mathbf{x}_i \; - \frac{1}{2} \bar{\mathbf{x}}_k' \boldsymbol{S}^{-1} \bar{\mathbf{x}}_k + \log(\pi_k) \tag{4.14}$$

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

$$L_k(\mathbf{x}_i) > L_l(\mathbf{x}_i) \tag{4.15}$$

for all $k \neq l$, where $L_k(\mathbf{x}_i)$ is given in equation (4.14). Classifying observations based on equation (4.15) is called linear discriminant analysis (LDA).

LDA and QDA are widely used and computationally efficient methods. LDA performs well if distributions are multivariate normal and group covariance matrices are identical. However, QDA performs well if distributions are multivariate normal and group covariance matrices are not identical. However, both LDA and QDA are not effective in handling data from nonnormal distributions. In the next two sections, two approaches to handle this problem will be explained.

4.3 *k*-Nearest neighbor discriminant analysis: *k*-NNDA

One of the popular approaches to handle the problem of nonnormal distributions is *k*-nearest neighbor discriminant analysis (*k*-NNDA). *k*-NNDA is the simplest machine learning algorithm, and it does not make any assumptions about the underlying probability distribution of the observations. The basic concept of *k*-NNDA is that it classifies an observation into a group which is the most common among its *k*-nearest

neighbors. For example, in Figure 4.1, we want to classify a green star into the blue circle group or yellow circle group. When $k = 1$, the green star is classified into the blue circle group which is the closest. When $k = 3$, the green star is classified into the red circle group because 2 among 3 nearest neighbors are red circle.

In *k*-NNDA, the posterior probability of observation x_i belonging to group *k* is given by:

$$P(k|x_i) = \frac{\pi_k m_k}{\sum_{k=1}^{K} \pi_k m_k} \tag{4.16}$$

where m_k represents the number of observation that are in the neighborhood of the $\mathbf{x}_i$ that belong to group *i*. The posterior probability, $P(k|x_i)$, of a given observation is proportional to m_k which is the number of units in the neighborhood of $\mathbf{x}_i$ belonging to group k. As illustrated in Figure 4.1, an observation is classified into a group where the posterior probability of group membership is the greatest.

The neighborhood of x_i is defined by the distance from x_i to the kth nearest neighbor (Huberty & Olejnik, 2006). Either the Euclidean distance or the Mahalanobis distance is usually used to calculate the distance. The Euclidean distance is defined by:

$$d = \sqrt{(x - x_i)'(x - x_i)} \tag{4.17}$$

And the Mahalanobis distance is given by:

$$d = \sqrt{(x - x_i)'\Sigma^{-1}(x - x_i)} \tag{4.18}$$

where Σ is the covariance matrix.

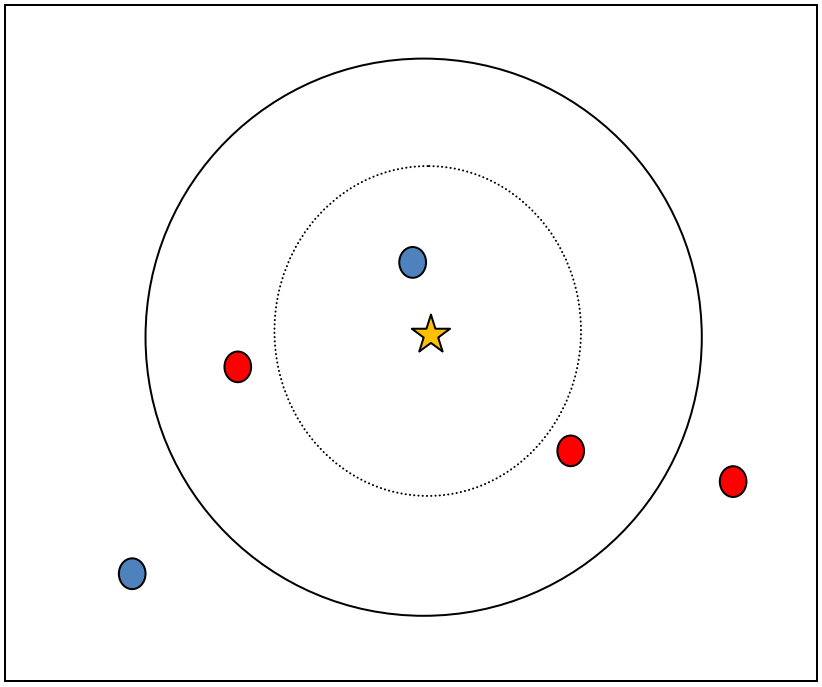


Figure 4.1 The classification rule of *k*-NNDA

4.4 Kernel density estimate discriminant analysis : KDEDA

4.4.1 Overview of KDEDA

KDEDA is another non-parametric approach to handle nonnormal probability distributions. It uses kernel density estimators instead of the normal density assumption for calculating the conditional probabilities. Kernel density estimation is a non-parametric density estimation approach which has no fixed data structure, and depends on all the data points to reach an estimate.

Suppose that $x_{k1}, x_{k2}, \ldots, x_{kn_k}$ are *p*-dimensional observations from the *k*-th population. Then, the multivariate kernel density estimator is given by

$$\hat{f}_k(\boldsymbol{x}) = {n_k}^{-1} h^{-p} \sum_{i=1}^{n_k} K\{h^{-1}(\boldsymbol{x} - \boldsymbol{x}_{ki})\} \tag{4.19}$$

where *K(•)* is a kernel function, and *h* is smoothing parameter known as bandwidth matrix. For our application, we will implement the multivariate Gaussian kernel, and focus on finding an optimal bandwidth matrix, because the performance of a kernel density estimator is primarily determined by the choice of bandwidth, and only in a minor way by the choice of kernel function (Zhang, King, & Hydman, 2004). The multivariate Gaussian kernel is given by

$$K(\boldsymbol{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2} \boldsymbol{x}^t \Sigma^{-1} \boldsymbol{x}\right\} \tag{4.20}$$

and the multivariate Gaussian kernel density estimator is given by

$$\widehat{f_k}(\boldsymbol{x}) = \frac{1}{n(2\pi)^{p/2}} \sum_{i=1}^{n_k} |H_k|^{-1/2} \exp\{(\boldsymbol{x} - \boldsymbol{x_i})' H_k{}^{-1} (\boldsymbol{x} - \boldsymbol{x_i})\} \qquad (4.21)$$

To classify observations into a group, we plug in the multivariate Gaussian kernel density estimate into the Bayes' rule. Then, discrimination rule becomes:

$$d_h(\boldsymbol{x}) = arg_k^{max}(\hat{\pi}_k \widehat{f_k}(\boldsymbol{x})) \qquad (4.22)$$

where $\hat{\pi}_k$ is the prior probability of the *k*-th group (*k=1,2,...K*). According to this rule, we assign $\boldsymbol{x}$ to the group *k* for which $\hat{\pi}_k \widehat{f_k}(\boldsymbol{x})$ is maximized.

4.4.2 The choice of optimal bandwidth matrix

As mentioned earlier, the correct choice of an optimal bandwidth matrix is a critical factor for the performance of the kernel density estimator. However, there are only a few papers published which discuss selecting the optimal bandwidth for the multivariate kernel (Zhang, King, & Hydman, 2004). This is primarily due to the computational difficulty in finding a data-adaptive optimal bandwidth matrix. Several approaches to find an optimal bandwidth matrix will be explained next.

One approach to find an optimal bandwidth matrix is to use cross-validation techniques to minimize the misclassification rate for different bandwidths. Sain, Baggerly and Scott (1994) compared the performance of the biased cross validation method, the least-squares cross-validation method, and bootstrap method for bandwidth selection in multivariate density estimation. They found that the biased cross-validation method performed well compared to other two methods. However, they also found that the

problem of selecting an optimal bandwidth matrix in kernel density estimation grows in complexity as the dimensionality of data increases. Additionally, cross-validation methods sometimes find multiple values of the bandwidths to minimize the misclassification rate, from which it is difficult to identify an optimal bandwidth (Ghosh & Bandyopadhyay, 2006).

Zhang, King and Hydman (2004) proposed using Markov chain Monte Carlo (MCMC) algorithms. They treated the elements of the bandwidth matrix as parameters whose posterior density can be obtained through the likelihood cross-validation criterion. They found that the MCMC algorithm generally performed better than the bivariate plug-in algorithm of Duong and Hazelton (2003) and the normal reference rule discussed in Bowman and Azzalini (1997). Yet, they also mentioned that the computation time for higher dimesional data did increase. Increased computational time for datasets wih high dimensionality make its application to discriminant analysis impractical.

Bozdogan (2007) presented eight different structures of the bandwidth matrix. Most of the bandwidth matrix structues are derived from estimating the bandwidth matrix based on the structure of the covariance matrix. The estimated bandwidth matrix can be given by:

$$\widehat{H} = n^{-1/(p+4)}\widehat{\Sigma}^{1/2} \tag{4.23}$$

where p is the dimension of data, and $\widehat{\Sigma}$ is the estimated covariance matrix. Table 4.1 provides seven different covariance structures which were compiled by Bensmail and Bozdogan (2002). In Table 4.1,

Table 4.1 Description of covariance structures

Model	Shape	Volume	MLE
1.λI	Spherical	Same	$\lambda = (np)^{-1}tr(W)$
2.$\lambda_k I$	Spherical	Different	$\lambda_k = (n_k p)^{-1}tr(W_k)$
3.B	≠Ellipsoidal	Same	$B = n^{-1}diag(W)$
4.λB_k	≠Ellipsoidal	Same	$\lambda = n^{-1}\sum_k \lvert diag(W_k)\rvert^{1/p}$ $B_k = \lvert diag(W_k)\rvert^{-\frac{1}{p}} diag(W_k)$
5.$\lambda_k B_k$	≠Ellipsoidal	Different	$\lambda_k = n_k^{-1}\lvert diag(W_k)\rvert^{1/p}$ $B_k = \lvert diag(W_k)\rvert^{-\frac{1}{p}} diag(W_k)$
6.Σ	Linear Kernel	Same	n^{-1}W
7.Σ_k	Linear Kernel	Different	$n_k^{-1}W_k$

$$W_k = \Sigma\,(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)' \text{ and}$$

$$W = \sum_{k=1}^{K} W_k \tag{4.24}$$

Another form of bandwidth matrix which was proposed by Bozdogan (2007) is based on the nearest neighbor. In this form, the bandwidth matrix for each group is a diagonal matrix

$$H = diag\,(h_1, \dots, h_K) \tag{4.25}$$

with

$$h_k = \left[\frac{1}{n_k p}\sum_{i=1}^{n_k} d^2(\mathbf{x}_i, 2 - NN_i)\right]^{1/2} \tag{4.26}$$

on the main diagonal.

The optimal bandwidth matrix among these 8 forms is selected based on Bozdogan's ICOMP. The general form of ICOMP can be defined by:

$$ICOMP = -2logL\big(x_{1,}x_{2,}\dots, x_n\big|H_k) + 2C_{1F}(\hat{\Sigma}) \tag{4.25}$$

For KDE bandwidth selection, the complexity part of ICOMP, $C_{1F}(\hat{\Sigma})$, becomes

$$C_{1F}\big(\hat{\Sigma}\big) = \frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda}_a)^2 \tag{4.26}$$

where $\bar{\lambda}_a$ is the arithmetic mean of the eigen-values of the covariance matrix, and *s* is the dimension of covariance matrix. We choose the bandwidth matrix which provides the minimum value of ICOMP as the optimal bandwidth matrix.

4.4.3 Numerical example of bandwidth matrix

Example 1. Wine data

The wine data set has n=178 observations and p=13 variables from three different classes. There are n_1=59 observations in group 1, n_2=71 observations in group 2, and n_3=48 observations in group 3.

For the purpose of illustration, only two variables are used to explain how to choose the optimal bandwidth matrix. For this data, variable X_6 and X_{10} are arbitrarily selected. X_6 represents phenol contents and X_{10} represents color intensity of wine. Figure 4.2 shows the contour plot and surface plot of the wine data. It suggests that each group does not have unique characteristics so that group membership of some observations are not clear in terms of variables X_6 and X_{10}.

The optimal covariance structure for group1, group 2 and group 3 which is chosen by the ICOMP value is B. As shown in Table 4.2, the covariance structure, B, has the minimum ICOMP value among the eight potential covariance structures. The calculated optimal bandwidth matrix for each group is as follows:

For group 1,

$$H_1 = \begin{bmatrix} 0.3817 & 0 \\ 0 & 1.3194 \end{bmatrix}$$

For group 2,

$$H_2 = \begin{bmatrix} 0.3374 & 0 \\ 0 & 1.1662 \end{bmatrix}$$

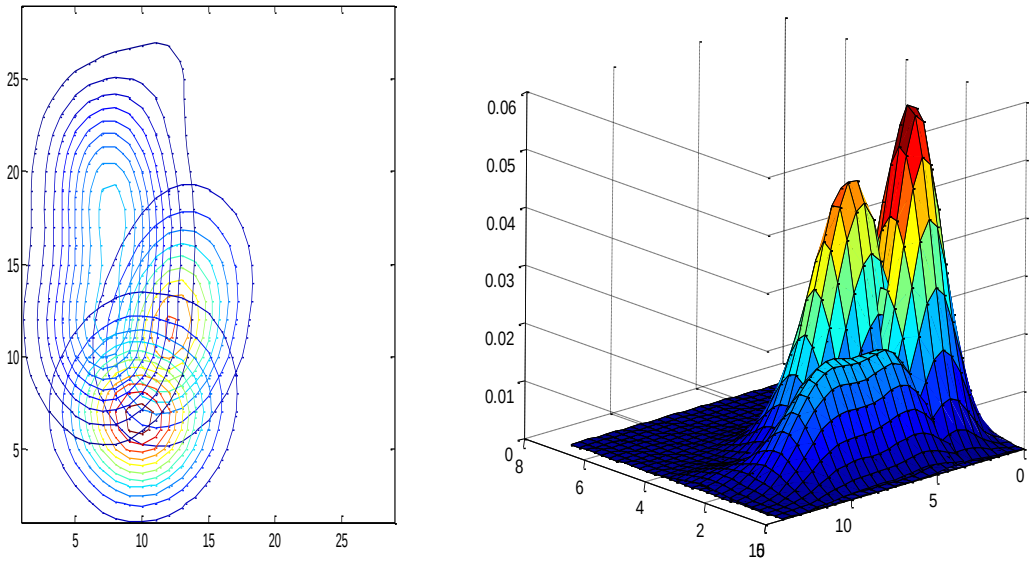


Figure 4.2 Contour and surface plots of the wine data for 2 variables

Table 4.2 ICOMP scores for potential covariance structures for the wine data

Group	λI	$\lambda_k I$	B	λB_k	$\lambda_k B_k$	Σ	Σ_k	NN
1	2836.1	2881.7	2747.5	2910.6	2978.6	2775.0	3111.4	2907.4
2	3471.4	3773.6	3081.3	3539.4	3590.2	3085.5	3622.3	3780.2
3	2900.3	2840.0	2812.7	3026.1	2965.2	2856.6	3011.8	2978.2

For group 3,

$$H_3 = \begin{bmatrix} 0.4380 & 0 \\ 0 & 1.5140 \end{bmatrix}$$

Based on these optimal bandwidth matrixes, we can classify each observation into one of three groups according to (4.21) and (4.22). The result of classification is given by Table 4.3. The overall classification error rate of the wine data in terms of variables X_6 and X_{10} is 0.1348.

Example 2. Iris data

The iris data set has n=150 observations and *p*=4 variables from three different groups. Each group has 50 observations. Again, for the purpose of illustration, only two variables are used to demonstrate how to choose the optimal bandwidth matrix. Variables, X_2 and X_3, which represent sepal width and petal length, were arbitrarily chosen.

Figure 4.3 shows the contour and surface plots of the iris data. It shows that group 1 is clearly different from other groups, but there is no unique difference between group 2 and group 3.

The optimal covariance structure for group1, group 2 and group 3 which is chosen by the ICOMP scores is Σ . As it is shown in Table 4.4, the covariance structure, Σ, has the minimum ICOMP scores among the eight potential covariance structures for each group. Therefore, the three groups have the same bandwidth matrix structures. The calculated optimal bandwidth matrices for all three groups are as follows:

Table 4.3 Confusion matrix of the wine data

		Classified group			Total
		1	2	3	
Actual group	1	51	5	3	59
	2	7	63	1	71
	3	8	8	32	48
Total		66	76	36	178

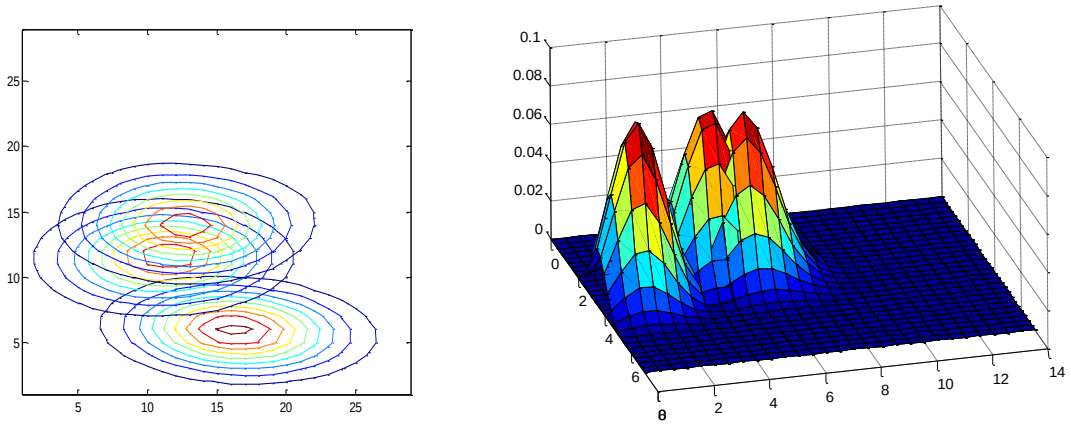


Figure 4.3 Contour and surface plots of the iris data for 2 variables

Table 4.4 ICOMP scores for potential covariance structures for the iris data

Group	λI	$\lambda_k I$	B	λB_k	$\lambda_k B_k$	Σ	Σ_k	NN
1	7024	8632	4617	9722	12982	2969	13458	4207
2	2898.1	2877.9	2483.6	2848.4	2809.7	2144.1	3178.2	2445.5
3	4573.7	4228.5	3359.1	4162.9	3875.1	2488.4	4219.5	3276.9

$$H_{1,2,3} = \begin{bmatrix} 0.2120 & -0.0787 \\ -0.0787 & 0.9103 \end{bmatrix}$$

Based on these optimal bandwidth matrices, we can classify each observation into one of three groups according to (4.21) and (4.22). The result of classification is given by Table 4.5. The overall classification error rate for the iris data in terms of variables X_6 and X_{10} is 0.08.

4.5 Model selection : New hybrid approach

There are several ways to choose the best model from several competing models. We could examine all possible models, or exploit various model search algorithms such as a stepwise method or tabu search. In this section, we will develop a new approach which will combine ICOMP and the GA for KDEDA. The performance of this new hybrid approach will be evaluated in Chapter 5.

Table 4.5 Confusion matrix of the iris data

		Classified group			Total
		1	2	3	
Actual group	1	50	0	0	50
	2	0	46	4	50
	3	0	8	42	50
Total		50	54	46	150

4.5.1 ICOMP for DA

Suppose we have a DA model with n observations and k predictor variables. It can be written in matrix form, which is given by:

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

$$\text{and}, \varepsilon_i = N(0, \sigma^2) \tag{4.27}$$

in more detail,

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{and } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where y is a response variable which represents group membership. In equation (4.27), we assume that the random errors are normally distributed with $E[\varepsilon|X] = 0, and$ $Var(\varepsilon_i) = \sigma^2$.

Log likelihood function

Equation (4.27) can be expressed as follows:

$$y|X \sim N(X\beta, \sigma^2 I) \tag{4.28}$$

We can write (4.28) in terms of the probability density function (pdf) as:

$$f(y|X; \beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}(y - X\beta)\frac{1}{\sigma^2}(y - X\beta)\right\} \tag{4.29}$$

Then the joint pdf of $y_1, \ldots, y_n$ is given by (Bozdogan, 2006):

$$L(\beta, \sigma^2) = f(y_1, \ldots, y_n|X; \beta, \sigma^2) = \prod_{I=1}^{n} f(y_i|X; \beta, \sigma^2)$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - E(y_i)^2\right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - E(\boldsymbol{y})\right)'\left(\boldsymbol{y} - E(\boldsymbol{y})\right\}\right.$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\beta)'(\boldsymbol{y} - X\beta)\right\} \quad (4.30)$$

The log likelihood function of (4.30) can be written as:

$$l(\beta, \sigma^2) = lnL(\beta, \sigma^2)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - X\beta)'(\boldsymbol{y} - X\beta) \quad (4.31)$$

Now, we can get the maximum likelihood estimators$\hat{\beta}$ and $\hat{\sigma}^2$ by differentiating $l(\beta, \sigma^2)$.

First, we differentiate (4.31) in terms of σ^2:

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\boldsymbol{y} - X\beta)'(\boldsymbol{y} - X\beta)}{2\sigma^4} \quad (4.32)$$

Since

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = 0 \quad (4.33)$$

The maximum likelihood estimator

$$\hat{\sigma}^2 = \frac{1}{n}((\boldsymbol{y} - X\beta)'(\boldsymbol{y} - X\beta) \quad (4.34)$$

When equation (4.34) is maximized in terms of β, it can be written as:

$$\hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{y} - X\hat{\beta}\right)'\left(\boldsymbol{y} - X\hat{\beta}\right) \quad (4.35)$$

In equation (4.27), the error term, $\boldsymbol{\varepsilon}$, is defined as the difference between the observed group membership and the predicted group membership:

$$\boldsymbol{\varepsilon} = \boldsymbol{y} - X\hat{\beta}$$

$$= \boldsymbol{y} - \hat{\boldsymbol{y}} \quad (4.36)$$

Therefore, (4.35) can be written as:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.37}$$

Finally the maximized log likelihood function can be given by:

$$l(\hat{\theta}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{n}{2} \tag{4.38}$$

ICOMP for KDEDA: model selection criteria

To choose the best model among several competing models for KDEDA, ICOMP (Bozdogan, 1988, 1990, 1994, 2000, 2009) is used. Among several forms of complexity, both $C_{1F}\left(\hat{\Sigma}(\hat{\theta})\right)$ and $C_1\left(\hat{\Sigma}(\hat{\theta})\right)$ are appropriate for KDEDA, because both measures of complexity are invariant under an orthogonal transformation. In this dissertation, $C_{1F}\left(\hat{\Sigma}(\hat{\theta})\right)$ is utilized for the purpose of illustration.

As introduced in Chapter 2, the general form of ICOMP is given by:

$$ICOMP = -2l(\hat{\theta}) + 2C_{1F}\left(\hat{\Sigma}(\hat{\theta})\right) \tag{4.39}$$

To derive ICOMP for KDEDA, we substitute (4.38), the maximized log likelihood function of DA into (4.39). Then $ICOMP_{KDEDA}$ can be written by:

$$\begin{aligned} ICOMP_{KDEDA} &= -2l(\hat{\theta}) + 2C_{1F}\left(\hat{\Sigma}(\hat{\theta})\right) \\ &= -2\left\{-\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{n}{2}\right\} + 2C_{1F}\left(\hat{\Sigma}(\hat{\theta})\right) \end{aligned}$$

$$= n\ln(2\pi) + n\ln(\hat{\sigma}^2) + n + 2\left\{\frac{s}{4}\frac{\frac{1}{s}tr\left(\hat{\Sigma}'\hat{\Sigma}\right) - \left(\frac{tr\left(\hat{\Sigma}\right)}{s}\right)^2}{\left(\frac{tr\left(\hat{\Sigma}^{-1}\right)}{s}\right)^2}\right\} \quad (4.40)$$

where

$$s = rank\left(\hat{\Sigma}\left(\hat{\theta}\right)\right)$$

and

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

As introduced in Chapter 2, $C_{1F}\left(\hat{\Sigma}\left(\hat{\theta}\right)\right)$ can be expressed in terms of eigenvalues. Then, (4.40) can be given by:

$$ICOMP_{KDEDA} = n\ln(2\pi) + n\ln(\hat{\sigma}^2) + n + 2\left\{\frac{1}{4\bar{\lambda}_a^2}\sum_{j=1}^{s}(\lambda_j - \bar{\lambda}_a)^2\right\} \quad (4.41)$$

4.5.2 New hybrid approach for KDEDA

In this section, we introduce the new hybrid approach for KDEDA. We combine the GA, KDE and ICOMP to choose the best model for DA. We use this approach to simultaneously find an optimal bandwidth matrix for KDE and the best model from several competing models.

Our proposed approach is to use the multivariate Gaussian kernel density estimator instead of the Gaussian. We calculate the probability of group membership of

each observation based on the multivariate Gaussian kernel density estimator. Then, we assign each observation to a group with the maximum posterior group membership.

The GA is used as the main search algorithm to find the best among several competing models. In our approach, the GA uses ICOMP as an objective function which guides evolution over generations. A model with the lowest ICOMP values is chosen as the best model from each generation. The GA uses crossover and mutation operators to find better models in a predetermined number of generations.

This hybrid approach consists of two stages. (1) It finds an optimal bandwidth matrix for KDE of the given subset model. Among the eight bandwidth matrices, the one with the minimum ICOMP value will be chosen for the specific model. (2) It finds the best model from several competing models by using the GA driven by ICOMP. The GA will identify a model with the minimum ICOMP value as the best model. The pseudo code for KDEDA with the GA is shown here:

Main function

% Required function : Kdeda.m and Bandwidth.m

% Initiate model selection parameters

Number of generation

Number of population

Crossover rate

Mutation rate

Elitism

% Input data

% Initialize population - start out with about half 1s

% Begin genetic algorithm

```
      for gencnt = 1:num_generns
```

% Compute objective function values

```
            for popcnt = 1:popul_size
                  [pop_fitness(popcnt) err(popcnt)] =
KDEDA(sample,training,                    xgroup,population(popcnt,:));
            end  % chromosomes loop
```

% Sort scores appropriately

% roulette selection – to mate offspring

% Mutation operation

% Crossover operation to create offspring

% Elitism to forward the best individual to new generation

```
      end
```
% generations loop

% End genetic algorithm

Kdeda function

```
function [ICOMP err]=kdeda(sample,training,xgroup,bin)
```

% Get data

% Calculate the prior probability

% Choose the optimal bandwidth for each group

```
      [H ICOMP c]=bandwidth(y,training,lambda,lambda2,W);
```

% Calculate the posterior probability of each observation

% Assign each observation into one of groups

% Compute error rate

% Compute ICOMP

```
ICOMP=n*log(2*pi)+n*log(err)+n+2*sumC;
```

Bandwidth function

% Caluculate 8 bandwidth matrix structures

% Calculate kernel density estimation

% Calculate the maximized log likelihood function

% Calculate the eigenvalues of bandwidth matrices

```
lam=eig(Hs{i,1});
```

% Compute complexity

```
C(i)=1/4*(1./lamhat.^2).*lsumsum;
```

% Score ICOMP for 8 bandwidth matrices

```
ICOMP = [ICOMP -2*ll+2*C(i)];
```

% Find smallest ICOMP for each group

% Return optimal bandwidth matrix for each group

Chapter 5

Applications & Numerical Examples

The purpose of this chapter is to demonstrate application of the new hybrid approach for KDEDA which combines KDE, the GA and ICOMP on several numerical examples. The performance of KDEDA is compared with that of LDA, QDA and *k*-NNDA by using four real data sets. The *classification error rate*, which is defined in terms of the proportion of observations classified incorrectly, is used to evaluate models. We separate our data sets into two groups, a training sample and a test sample. The classification error rate from the test sample is used to evaluate several competing models and to compare different DA methods.

The material here is divided into 4 sections. Section 5.1 applies our proposed approach to the Iris data set, Section 5.2 applies it to the aorta data set, Section 5.3 applies it to the French data set, and Section 5.4 applies it to the college data set.

5.1 Iris data

5.1.1 Description of data set

This data set is from Sir Ronald Aylmer Fisher (1936) as an example of discriminant analysis. The data set consists of 50 observations from each of three different species of iris flowers (iris setosa, iris virginica and iris versicolor). It has four

variables which measure the sepal and petal lengths and widths. X_1 is the sepal length, X_2 is the sepal width, and X_3 and X_4 are the petal length and width. Prior probabilities are calculated based on the number of observations in each group. For validation, 75% of observations are partitioned into the training sample and 25% of observations are saved as the test sample.

In Figure 5.1, scatter plots of the iris data are provided. Observations from iris setosa are depicted by the "red circle," observations from iris virginica are depicted by the "pink triangle," and observations from iris versicolor are represented by the "blue square." These scatter plots suggest the three groups are separated with their own means. If we pay more attention to these scatter plots, however, we can identify that some observations from iris virginica are overlapped with observations from iris versicolor. To correctly assign these overlapped observations, we need to pay close attention to selecting appropriate models.

5.1.2 The result of GAs

The purpose of this data set is to determine the species for each observation, based on the length and width of the sepals and petals. For this data set, we may not need to use the GA, because it is easy to explore all the possible solutions – there are only $(\mathbf{2^4 - 1})$=15 possible solutions. However, the GA is convenient to use and it will not take significant computational time for this small data set.

The GA parameters are given in Table 5.1. We only performed 5 generations with

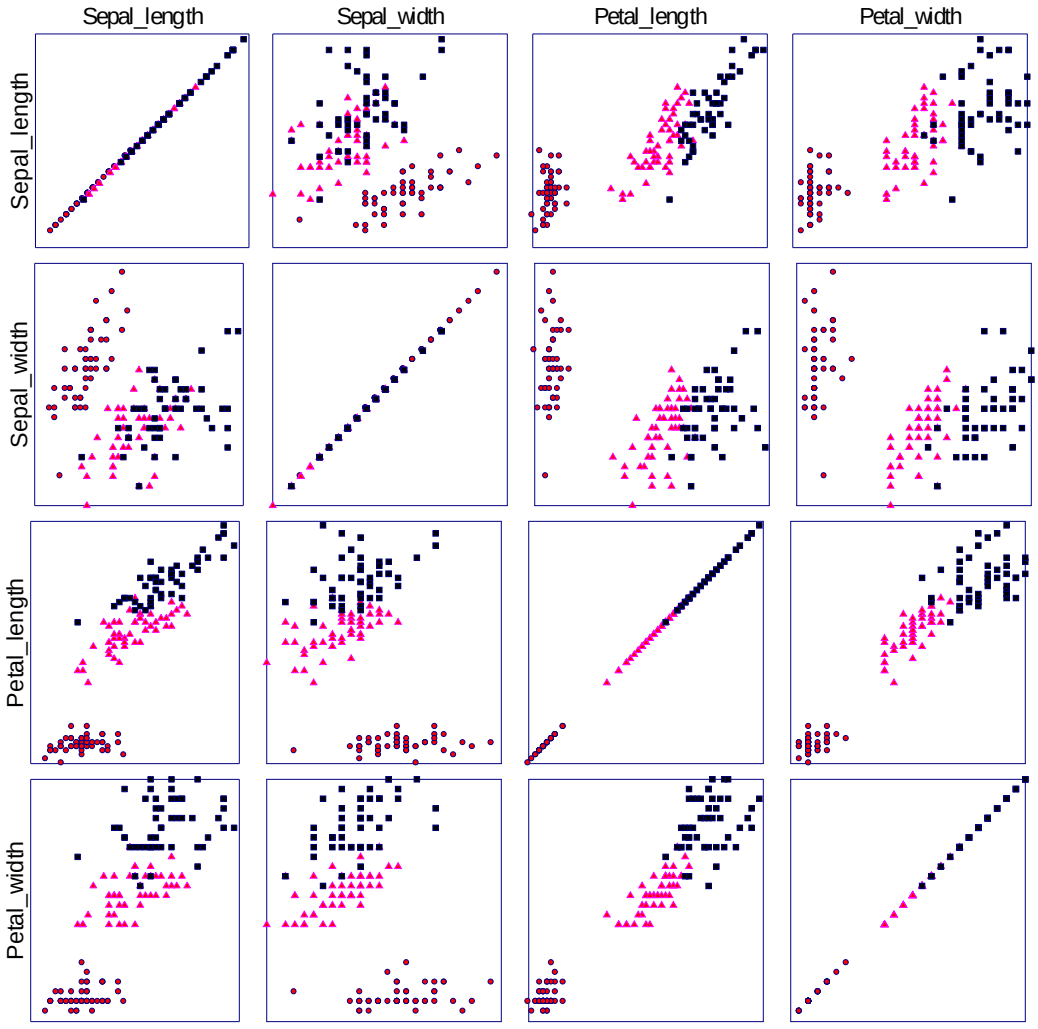


Figure 5.1 Scatter plots of the iris data (Circle=setosa, Triangle=virginica, Square= versicolor)

Table 5.1 GA parameters of the iris data example

Parameter	Value
Size of population	5
Number of generation	5
Fitness value	$C_{1F}\left(\hat{\Sigma}(\hat{\theta})\right)$
Probability of crossover	0.75
Probability of mutation	0.10
Elitism	Yes

5 individuals in each population, exploring at most 25 (possibly non-unique) models. The small number of generations and population size significantly reduces computational time. The result of one run of the GA is given in Figure 5.2 It only took 84 seconds, and, after three generations, it found the model with variables X_3, X_4 (ICOMP=-54.70) as the best model. In this case, all 10 GA replications found the model with variables X_3, X_4 as the best model. This model has 3.57% probability of misclassification for the training sample, and 0% probability of misclassification for the test sample.

5.1.3 Comparison of KDEDA with LDA, QDA and k-NNDA

In this section, we compare the performance of KDEDA with LDA, QDA and *k*-NNDA. For this data set, we analyzed all the possible models. This allows us to compare the performance of KDEDA with LDA, QDA and *k*-NNDA for several competing

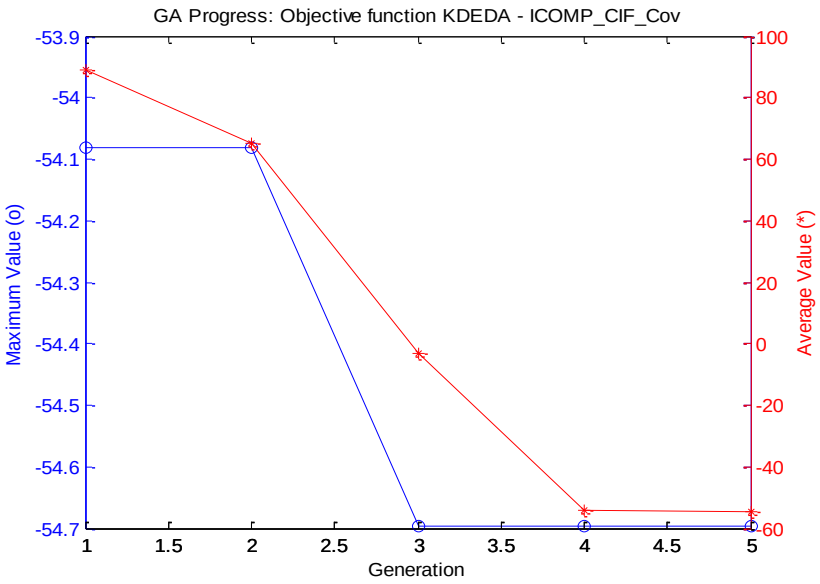


Figure 5.2 One run of the GA for the iris data

models. In addition, it also allows us to confirm whether or not the GA identified the best model for KDEDA.

In Table 5.2, we report the result of the four different DA methods for all possible solutions. Based on the ICOMP value of KDEDA, the model composed of variables X_3, X_4 (ICOMP = -54.70) is the best model, confirming the GA's selection. This model has 0 % probability of misclassification for the test sample for KDEDA, LDA and QDA, respectively, and 5.26% probability of misclassification for *k*-NNDA. The model with variables X_2, X_3, X_4 (ICOMP = -54.08) was chosen as the second best model based on the ICOMP value for KDEDA. In this case, the probability of misclassification for the test sample for KDEDA, LDA was 0%, and the probability of misclassification for QDA and *k*-NNDA were 5.26%, respectively. There were indistinguishable differences in ICOMP values for both models (-54.70 vs. -54.08). Both models can be regarded as the best model, and this is supported by the performance of the four DA methods. Based on the classification error rates of the iris data, our proposed approach performed as well as or better than the other three DA methods.

5.2 Aorta data

5.2.1 Description of data set

Our next data set is nuclear magnetic resonance (NMR) aorta imaging data from a study of heart disease, collected by Pearlman (1986) at the Medical school of the University of Virginia. The data set consists of 418 observations from 20 image

Table 5.2 Classification error rates of the iris data for different DA methods

Model	KDEDA				*k*-NNDA		LDA		QDA	
	ICOMP	Band Type	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample
X_1	184.32	3/3/3	0.3036	0.2632	0.2589	0.4211	0.2679	0.2105	0.2946	0.2105
X_2	227.52	7/4/7	0.4464	0.4474	0.4375	0.4737	0.4464	0.4474	0.4196	0.3947
X_3	22.27	6/6/6	0.0714	0	0.0357	0.0789	0.0446	0.0526	0.0446	0.0526
X_4	-9.95	6/6/6	0.0536	0	0.0536	0	0.0536	0	0.0536	0
X_1, X_2	154.37	3/6/6	0.2321	0.1842	0.1339	0.2632	0.2054	0.1842	0.2411	0.2368
X_1, X_3	35.90	6/6/6	0.0804	0	0.0446	0.0263	0.0357	0.0526	0.0446	0.0263
X_1, X_4	22.54	6/6/6	0.0714	0	0.0446	0.0526	0.0536	0	0.0446	0
X_2, X_3	47.67	6/6/6	0.0714	0.0263	0.0268	0.0526	0.0536	0.0263	0.0625	0.0526
X_2, X_4	22.34	8/6/6	0.0714	0.0263	0.0536	0.0263	0.0536	0	0.0536	0.0526
X_3, X_4	-54.70	6/6/6	0.0357	0	0.0179	0.0526	0.0625	0	0.0357	0
X_1, X_2, X_3	48.31	6/6/6	0.0893	0	0.0357	0.0263	0.0357	0.0526	0.0446	0.0526
X_1, X_2, X_4	22.27	6/6/6	0.0714	0.0263	0.0446	0.0526	0.0446	0.0263	0.0446	0.0526
X_1, X_3, X_4	23.61	6/6/6	0.0714	0	0.0268	0.0263	0.0268	0	0.0268	0
X_2, X_3, X_4	-54.08	6/6/6	0.0357	0	0.0357	0.0526	0.0357	0	0.0179	0.0526
X_1, X_2, X_3, X_4	-7.85	6/6/6	0.0536	0	0.0268	0.0263	0.0268	0	0.0268	0.0526

acquisition and direction and orientation variables. The first group of 194 patients exhibited early atheroma, and the second group of 224 patients were healthy. The prior probabilities for two groups are 46.4% for group 1, and 53.6% for group 2. For validation, 70% (125 obs.) of the observations are partitioned into the training sample and 30% (53 obs.) of the observations are saved as the test sample.

In Figure 5.3, selected scatter plots in terms of variables X_6, X_7, X_8, X_9 are provided. Observations in group 1 are depicted by the "red circle", and observations in group 2 are depicted by the "black triangle". These scatter plots suggest two groups are separated with their own means. If we pay more attention to these scatter plots, however, we can identify that some observations in one group are overlapped with observations in the other group. To correctly assign these overlapped observations, we need to pay close attention to selecting appropriate models.

5.2.2 The result of GAs

This data set has $(2^{20} - 1) = 1{,}048{,}575$ possible solutions. We performed 20 replications of the GA with 30 individuals and 20 generations. Thus one run of the GA analyzed at most 600 models, accounting for only 0.06% of all possible models. Other GA parameters are given in Table 5.3

The typical example of one run of the GA is given in Figure 5.4 and Table 5.4. It required 5.5 hours (334 minutes) to finish computation, and found the model $\{X_1\}$, as the best model, with the minimum ICOMP value of -3677.51. In Table 5.4, we can recognize that the GA identified more parsimonious models over generations. The number of

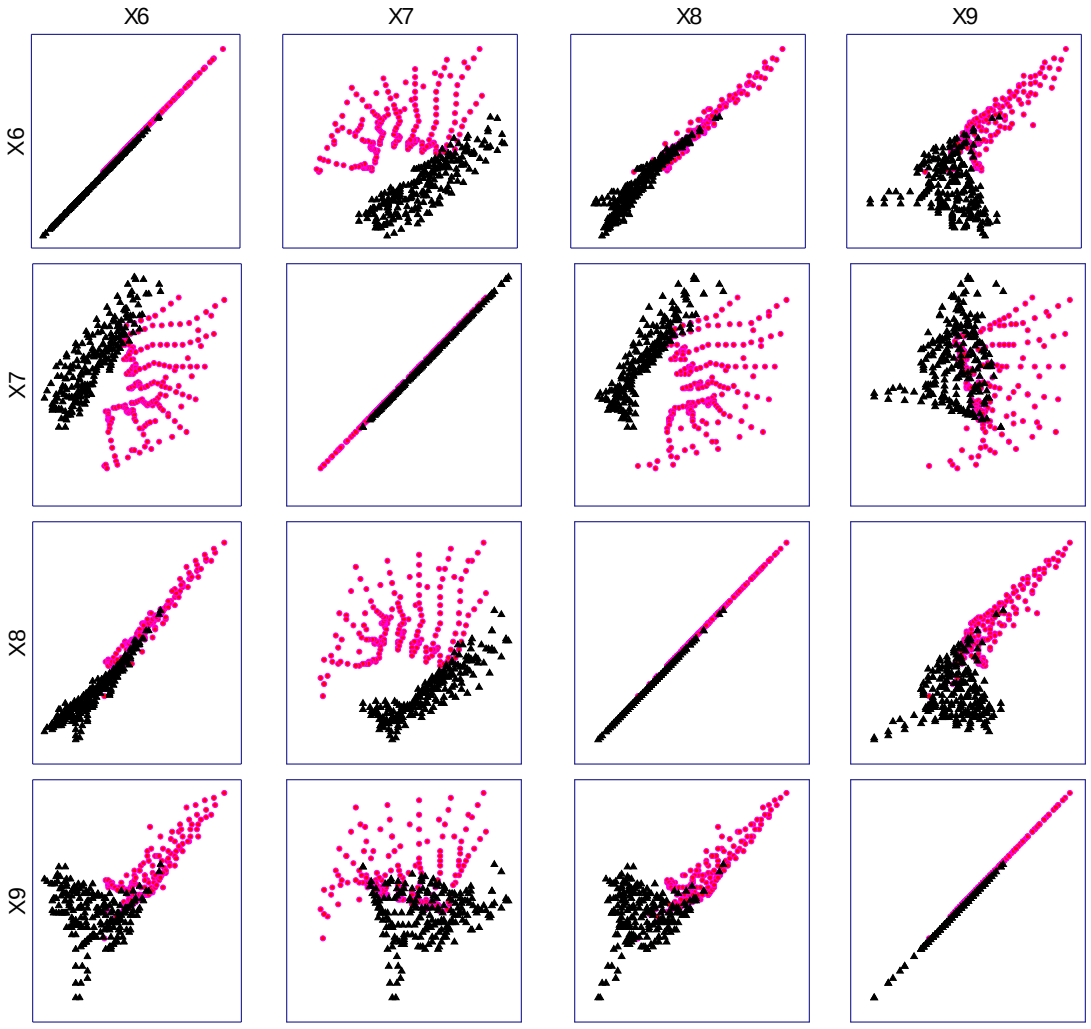


Figure 5.3 Scatter plots of the aorta data

Table 5.3 GA parameters of the aorta data

Parameter	Value
Size of population	30
Number of generation	20
Fitness value	$C_{1F}\left(\hat{\Sigma}(\hat{\theta})\right)$
Probability of crossover	0.75
Probability of mutation	0.10
Elitism	Yes

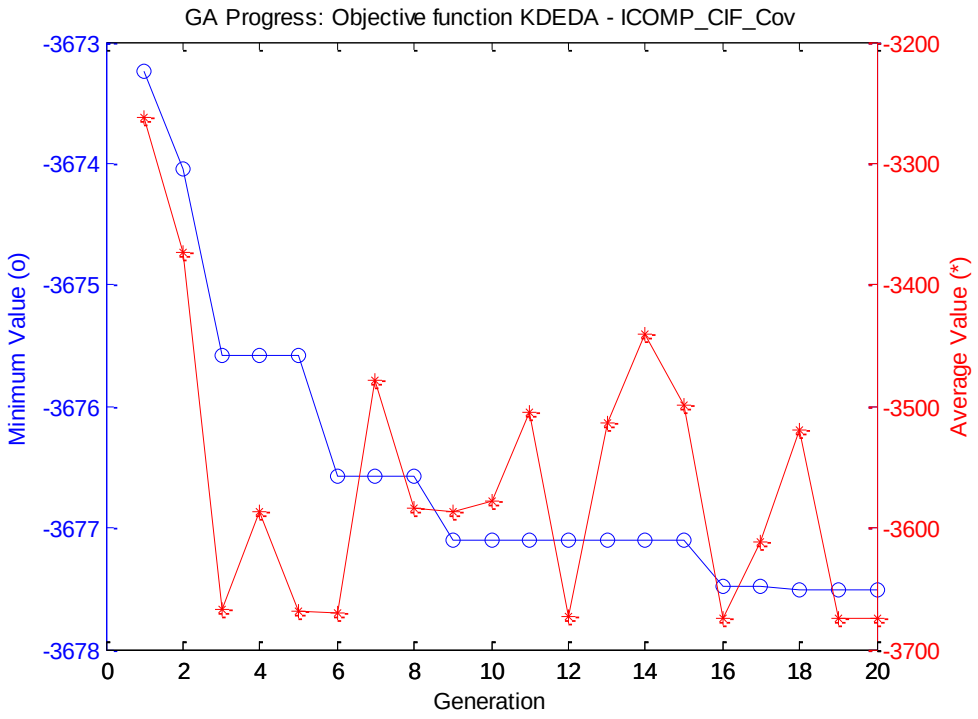


Figure 5.4 Example of a run of the GA for the aorta data

Table 5.4 Models selected by one run of the GA for the aorta data

Subset	ICOMP	Frequency	Training sample error rate
X_1	-3677.51	3	0
X_1, X_{15}	-3677.48	2	0
X_1, X_{13}	-3677.10	7	0
X_4, X_{13}	-3676.57	3	0
X_2, X_5, X_{14}, X_{16}	-3675.59	3	0
$X_2, X_5, X_{11}, X_{18}, X_{20}$	-3674.04	1	0
$X_2, X_5, X_{11}, X_{13}, X_{15}$	-3673.24	1	0

variables in a model decreased from 5 to 1. This result satisfies an objective of model selection algorithms, which should find as simple model as possible.

The best solutions chosen across 20 replications of the GA are shown in Table 5.5. It seems reasonable to have 15 models chosen as the potential best models for 20 runs of the GA, because one replication only searched 0.06% of all the possible models. The model,$\{X_1\}$, had the minimum ICOMP value of -3677.51, and the model $\{X_1, X_{15}\}$, had the ICOMP value of -3677.48. Though the two models have similar ICOMP values, the principle of parsimony tells us to regard $\{X_1\}$ as the best model. This model was selected 2 times over 20 replications of the GA. This result suggests that our proposed approach did not demonstrate very good performance in finding the best model – possibly a consequence of the enormous number of possible models. If we increase the size of population and the number of generation for the GA, our approach will be able to show better consistency in finding the best model.

5.2.3 Comparison of KDEDA with LDA, QDA and k-NNDA

In this section, we again compared the performance of KDEDA with LDA, QDA and *k*-NNDA, by analyzing the 15 models selected by the GA. In Table 5.6, we report the classification error rates of selected models for different DA methods.

First, in terms of the best model, $\{X_1\}$, all methods including KDEDA, *k*-NNDA, LDA, and QDA had 0% classification error rates for the test sample. The best model, $\{X_1\}$, chosen by ICOMP, showed excellent performance for all four methods.

Table 5.5 Models selected across 20 replications of the GA for the aorta data

Subset	ICOMP	Frequency
X_1	-3677.51	2
X_1, X_{15}	-3677.48	1
X_4, X_{19}	-3677.25	2
X_4, X_{20}	-3677.21	1
X_1, X_7	-3677.15	2
X_2, X_{12}	-3677.11	2
X_2, X_{12}, X_{14}	-3677.02	1
X_2, X_9, X_{14}	-3676.93	1
X_2, X_7, X_{17}	-3676.88	1
X_6, X_7, X_9	-3676.85	1
X_1, X_7, X_{17}	-3676.82	1
X_1, X_{10}, X_{13}	-3676.79	1
X_4, X_{15}	-3676.63	1
X_7, X_8, X_9	-3676.62	1
X_4, X_9	-3676.60	2

Table 5.6 Classification error rates of the aorta data for different DA methods

Variable	KDEDA		*k*-NNDA		LDA		QDA	
	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample
X_1	0	0	0	0	0	0	0	0
X_1, X_{15}	0	0	0	0	0	0	0	0
X_4, X_{19}	0	0	0	0	0.2987	0.2289	0.1219	0.0482
X_4, X_{20}	0	0	0	0	0.3911	0.3976	0.0313	0.0241
X_1, X_7	0	0	0	0	0	0	0	0
X_2, X_{12}	0	0	0	0	0.0886	0.0723	0	0
X_2, X_{12}, X_{14}	0	0	0	0	0.0829	0.0482	0.0094	0.0120
X_2, X_9, X_{14}	0	0	0	0	0.0571	0.0361	0.0031	0
X_2, X_7, X_{17}	0	0	0	0	0.0156	0	0	0
X_6, X_7, X_9	0	0.0120	0	0	0.0594	0.0482	0.0299	0.0241
X_1, X_7, X_{17}	0	0	0	0	0	0	0	0
X_1, X_{10}, X_{13}	0	0	0	0	0	0	0	0
X_4, X_{15}	0	0	0	0	0.4242	0.4699	0.0063	0
X_7, X_8, X_9	0	0	0	0	0.0529	0.0361	0.0330	0.0241
X_4, X_9	0	0	0	0	0.2617	0.2651	0.0187	0.0120
Mean	0	0.0008	0	0	0.1155	0.1068	0.0169	0.0096

Second, in terms of all 15 models, KDEDA and *k*-NNDA showed better performance compared to LDA and QDA. KDEDA had 0.08% classification error rate, *k*-NNDA had 0%, LDA had 10.68%, and QDA had 0.96%, respectively.

5.3 French data

5.3.1 Description of data set

Our third data set is about enrollment for college French classes, used by Huberty (1994) and Glen (2001). In this data, two groups of students are classified by enrollment for college French at beginner (group 1) and intermediate (group 2) levels. This data set consists of 13 characteristics for each student, which are shown in Table 5.7. There are n_1=35 observations in group 1 and n_2=81 observations in group 2. Therefore, prior probabilities for the two groups are 30.17%, 69.83%, respectively. This data set has a relatively small sample size. For validation, 80% (93 obs.) of observations are partitioned into the training sample and 20% (23 obs.) of observations are saved as the test sample.

In Figure 5.5, we show selected scatter plots of this data. These scatter plots show that many observations in different groups are overlapped with each other. This high proportion of overlapped observations will lower classification accuracy.

5.3.2 The result of GAs

This data set has $(2^{13} - 1) = 8191$ possible solutions. We performed 20 replications of the GA with 20 individuals and 20 generations. Thus one run of the GA

Table 5.7 The French data description

Variable	
X_1	Grade point averages in English
X_2	Grade point averages in mathematics
X_3	Grade point averages in social studies
X_4	Grade point averages in natural science
X_5	Number of semesters of high school French
X_6	Grade point averages in French
X_7	Aptitudes measures for English
X_8	Aptitudes measures for mathematics
X_9	Aptitudes measures for social studies
X_{10}	Aptitudes measures for natural sciences
X_{11}	French test scores in aural comprehension
X_{12}	French test scores in grammar
X_{13}	Number of semesters since the last high school French course

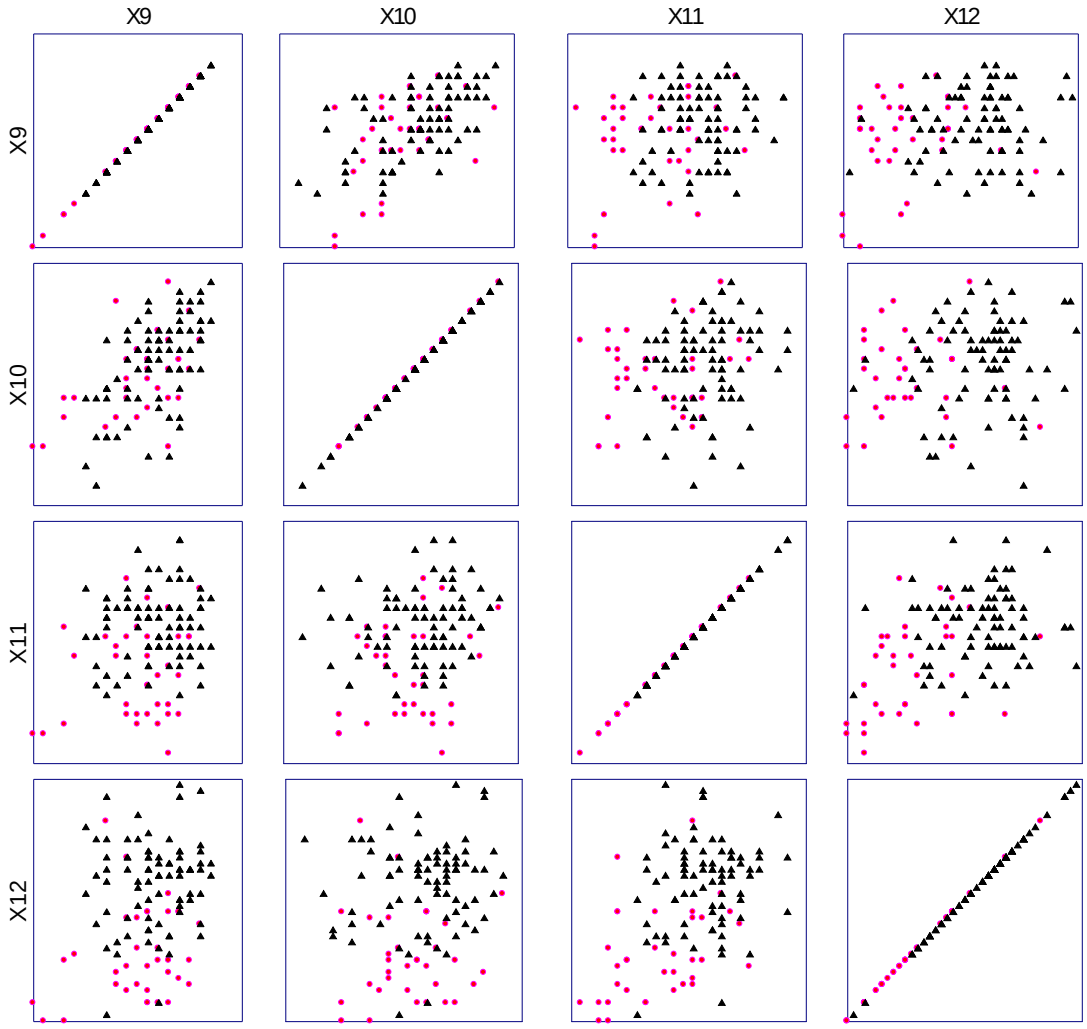


Figure 5.5 Selected scatter plots of the French data

analyzed at most 400 models, which accounted for only 4.88% of all the possible models. The remaining GA parameters are the same as those used for the aorta data set. The typical GA run for this data set took about 7 minutes (430 seconds).

The best solutions chosen across 20 replications of the GA are shown in Table 5.8. Nine models are selected across 20 GA replications. The model,$\{X_8, X_9, X_{11}, X_{13}\}$, had the minimum ICOMP value of -1020.10, and the model $\{X_3, X_4, X_6\}$, had the ICOMP value of -1020.04. These two models have similar ICOMP values, but in terms of the principle of parsimony, the model,$\{X_3, X_4, X_6\}$, can be regarded as the best model. This model was also the most frequently selected model, which was chosen 7 times among 20 replications. If we consider the number of possible models, 8191, this result satisfies our research question - whether the new hybrid approach incorporating KDEDA with the GA using ICOMP is compatible with the all-possible-subset approach.

5.3.3 Comparison of KDEDA with LDA, QDA and k-NNDA

In this section, we again compared the performance of KDEDA with LDA, QDA and *k*-NNDA. For this data set, we analyzed the 9 models selected by the GA by implementing LDA, QDA and *k*-NNDA.

In Table 5.9, we report the classification error rates of the four different DA methods for the selected models. In the case of the best model, $\{X_3, X_4, X_6\}$, KDEDA had a test sample error rate of 21.74% , *k*-NNDA had 13.04%, LDA had 26.09%, and QDA had 26.09%, respectively. *k*-NNDA showed the best performance for this data set, and KDEDA performed better than LDA and QDA. In terms of the average classification

Table 5.8 Models selected across 20 replications of the GA for the French data

Subset	ICOMP	Frequency
X_8, X_9, X_{11}, X_{13}	-1020.10	3
X_3, X_4, X_6	-1020.04	7
X_3, X_4, X_8	-1019.32	3
X_3, X_4, X_{10}	-1019.32	2
X_8, X_9, X_{11}, X_{12}	-1019.32	1
X_1, X_4, X_{10}	-1019.31	1
X_1, X_4, X_9	-1019.30	1
X_7, X_8, X_{11}, X_{12}	-1019.16	1
$X_9, X_{10}, X_{11}, X_{12}$	-1018.97	1

Table 5.9 Classification error rates of the French data for different DA methods

Variable	KDEDA		*k*-NNDA		LDA		QDA	
	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample
X_8, X_9, X_{11}, X_{13}	0	0.3043	0	0.1739	0.2294	0.3913	0.2468	0.3478
X_3, X_4, X_6	0	0.2174	0	0.1304	0.3008	0.2609	0.2849	0.2609
X_3, X_4, X_8	0	0.3913	0	0.2609	0.3341	0.2174	0.3095	0.3478
X_3, X_4, X_{10}	0	0.2609	0	0.2609	0.3579	0.3478	0.3405	0.3043
X_8, X_9, X_{11}, X_{12}	0	0.2609	0	0.1304	0.1802	0.1304	0.1730	0.1304
X_1, X_4, X_{10}	0	0.4348	0	0.3043	0.2746	0.3043	0.2587	0.3478
X_1, X_4, X_9	0	0.3913	0	0.2609	0.3000	0.4348	0.3413	0.4348
X_7, X_8, X_{11}, X_{12}	0	0.2609	0	0.1739	0.1802	0.1304	0.1389	0.1304
$X_9, X_{10}, X_{11}, X_{12}$	0	0.2609	0	0.2609	0.2048	0.1739	0.1881	0.1739
Mean		0.3092		0.2174		0.2657		0.2753

error rate, KDEDA had 30.92% classification error for the test sample, *k*-NNDA had 21.74% error rate, LDA had 26.57% error rate, and QDA had 27.53% error rate. These results suggest that other DA methods fit this data set better than KDEDA.

5.4 College data

5.4.1 Description of data set

The final data set is regarding college selectivity, provided by U.S News and World Report (2008). In this data set, colleges and universities are organized by how selective they are: that is, how picky they can be in choosing freshmen. Selectivity is determined by the test scores and high school class standing of applicants, plus the proportion of applicants who are accepted. These 9 variables are shown in Table 5.10.

Originally, this data set had 139 observations, but we omitted 16 observations with missing variables. Therefore, the total number of observations used here is 123. Among them, 34 colleges and universities are categorized into group 1 - most selective schools. The other 89 colleges and universities are categorized into group 2 - more selective schools. Therefore, the prior probability for group one is 27.64% and the prior probability for group two is 72.36%. For the purpose of validation, 80% of observations are partitioned into the training sample and 20% of observations are saved as the test sample.

In Figure 5.6, we show selected scatter plots of this data set. These scatter plots suggest two groups have their own means, and most of the observations are well

Table 5.10 The College data description

Variable	
X_1	Acceptance rate of applicants
X_2	SAT critical reading, 25th percentile
X_3	SAT critical reading, 75th percentile
X_4	SAT math, 25th percentile
X_5	SAT math, 75th percentile
X_6	ACT composite, 25th percentile
X_7	ACT composite, 75th percentile
X_8	Percentage of students who were in top 10% at high school class standing
X_9	Percentage of students who were in top 25% at high school class standing

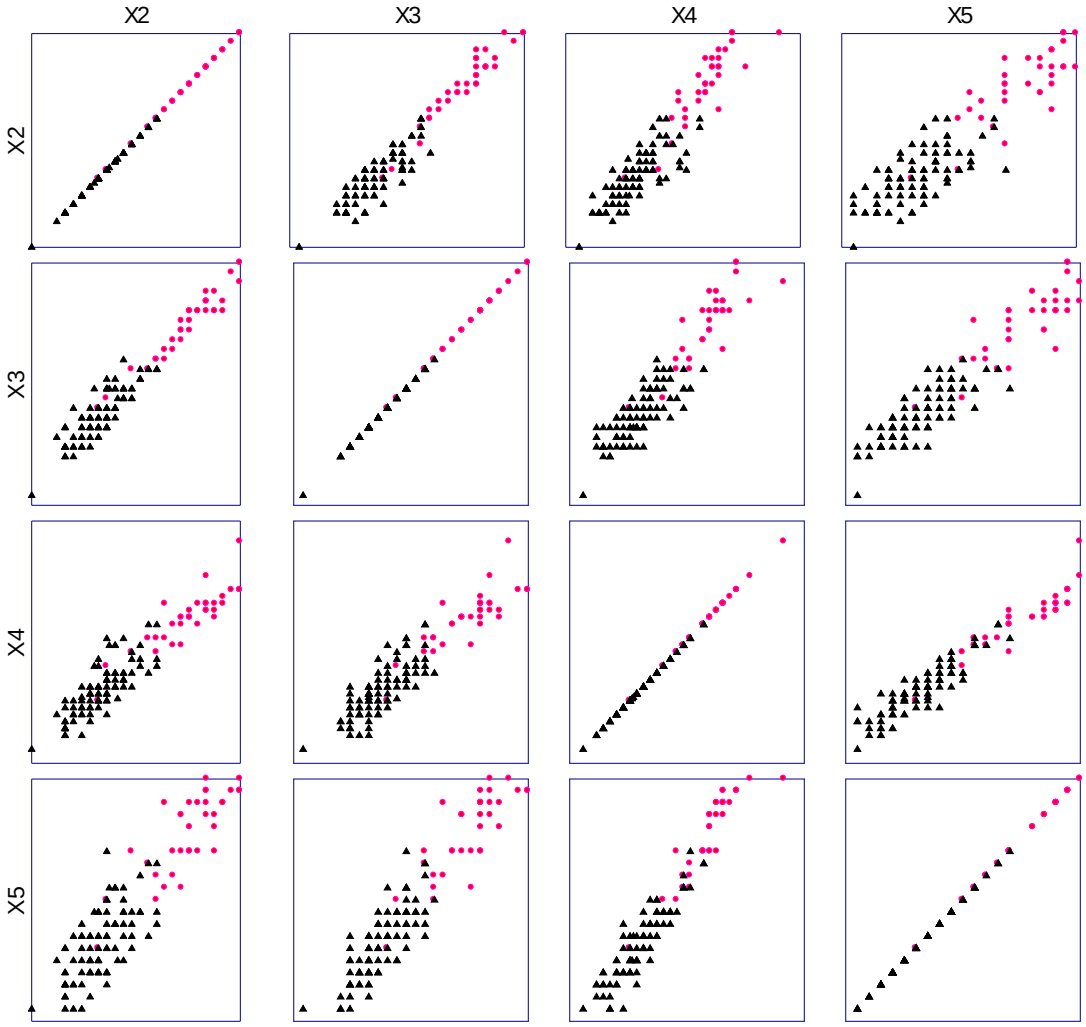


Figure 5.6 Selected scatter plots of the college data

separated from the other group. However, some observations in each group are overlapped. Therefore, it is necessary to select the best model maximizing classification accuracy.

5.4.2 The result of GAs

This data set has $(2^9 - 1) = 511$ possible solutions. We performed 20 replications of the GA with 20 individuals and 20 generations. Thus one run of the GA analyzed at most 400 models - 78.28% of the possible models. The typical GA run for this data set took about 12 minutes (710 seconds).

The best solutions chosen across 20 replications of the GA are shown in Table 5.11. Minimizing ICOMP, with a score of 1086.09, chose a model with variables, $\{X_3, X_8\}$, as the best; this model was also the most frequently selected model. It was chosen in 16 of the 20 GA replications. Again, our proposed approach showed consistency in finding the best model, suggesting that it is compatible with the all-possible-subset selection approach.

Table 5.11 Models selected across 20 replications of the GA for the college data

Subset	ICOMP	Frequency
X_3, X_8	-1086.09	16
X_2, X_8	-1086.07	4

5.4.3 Comparison of KDEDA with LDA, QDA and k-NNDA

In Table 5.12, we reported the classification error rate of all four different DA methods for selected models. In the case of the best model, $\{X_3, X_8\}$, KDEDA misclassified 4.17% of the test samples. Misclassification rates for the other methods were: *k*-NNDA had 4.17%, LDA had 8.33%, and QDA had 4.17%. KDEDA performed better than LDA, and performed as well as *k*-NNDA and QDA. In terms of the average of classification error rates, KDEDA, *k*-NNDA, and QDA had a 2.09% classification error rate for the test sample, and LDA had a 3.13% error rate. Again, KDEDA performed better than LDA, and performed as well as *k*-NNDA and QDA.

5.5 Conclusion

In this chapter, we applied our proposed approach to four real data sets to answer following two research questions: (1) whether KDEDA is superior to other methods: LDA, QDA, and *k*-NNDA, (2) whether the new hybrid approach incorporating

Table 5.12 Classification error rates of the college data for different DA methods

Variable	KDEDA		*k*-NNDA		LDA		QDA	
	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample	Training sample	Test sample
X_3, X_8	0	0.0417	0	0.0417	0.0208	0.0833	0.0347	0.0417
X_2, X_8	0	0	0	0	0.0208	0	0.0566	0
Mean		0.0209		0.0209		0.0313		0.0209

KDEDA with the GA using ICOMP is compatible with all-possible-subset solution.

Regarding the first research question, KDEDA showed better performance compared to LDA and QDA, and performed as well as *k*-NNDA. We show the classification error rates of the best model for each data set in Table 5.13. KDEDA and QDA performed the best on the iris data set, the aorta data set, and the college data set, but KDEDA exhibited less classification error than QDA for the French data set. *k*-NNDA had the least classification error rates for the aorta data, the French data, and the college data. LDA had the least classification error rates for the iris data set, and the aorta data set. In summary, KDEDA and *k*-NNDA showed better performance than LDA and QDA, based on these results.

In Table 5.14, we show the mean classification error rates of the test samples for each data set. KDEDA had the least mean classification error for the iris data, the aorta data, and the college data. *k*-NNDA showed the least mean classification error for the aorta data, the French data, and the college data. QDA had the lowest mean classification error rate for the college data. Again, KDEDA and *k*-NNDA showed better performance compared to LDA and QDA in terms of the mean classification error rate.

Regarding the second research question, our new hybrid approach may be compatible with the all-possible-subset method. Our proposed approach always found the best model for the Iris data which only had four variables and identified the best model in 16 out of 20 replications for the college data. Our proposed approach seems not to be as successful at finding the best model for the aorta data and the French data.

Table 5.13 Classification error rates for the best model of each data

Data set	Best model	KDEDA	*k*-NNDA	LDA	QDA
Iris data	X_3, X_4	0	0.0526	0	0
Aorta data	X_1	0	0	0	0
French data	X_3, X_4, X_6	0.2174	0.1304	0.2609	0.2609
College data	X_3, X_8	0.0417	0.0417	0.0833	0.0417

Table 5.14 Mean classification error rates of the test sample for each data set

Data set	KDEDA	*k*-NNDA	LDA	QDA
Iris data	0.0649	0.1088	0.0702	0.0824
Aorta data	0.0008	0	0.1068	0.0096
French data	0.3092	0.2174	0.2657	0.2753
College data	0.0209	0.0209	0.0313	0.0209

However, for these latter two datasets, the range of values of ICOMP for models which were chosen was less than one except for one model among 40 models. We may consider any of these models as equivalent to the best model in terms of ICOMP value. For this reason, we may conclude that our new hybrid approach can be compatible with the all-possible-subset selection method.

Chapter 6

Conclusion

6.1 Summary and conclusion

The purpose of this dissertation is to present a new approach which incorporates ICOMP, the GA, and KDE to handle both nonnormal distributions and high-dimensional data in the area of DA. We first introduced four different methods in DA. LDA and QDA are popular and widely used, but these are not effective when each group does not follow the Gaussian distribution. *k*-NNDA and KDEDA are nonparametric DA methods and they are used to handle the problem of nonnormal distributions. Then, we introduced Bozdogan's information-theoretic measure of complexity called ICOMP as a model selection criterion. ICOMP is based on the generalization of the covariance complexity index and was motivated in part by AIC. ICOMP shows better performance than AIC-type criteria and it has been applied into multivariate nonnormal regression models, threshold autoregressive models, neural networks, support vector machines, and so on. Finally, we introduced the genetic algorithm which is based on principles of natural selection to find an optimal solution.

In this work, we proposed a new hybrid approach for KDEDA. This is the most

significant contribution of this dissertation. For this, we derived the expression of ICOMP for KDEDA, which we used to drive the GA for KDEDA subset modeling. We use ICOMP as the objective function for the GA, and the GA identifies a model with the minimum ICOMP value as the best model. This approach enables researchers to find both an optimal bandwidth matrix for KDE and the best model from several competing models, which was a severe obstacle for researchers wishing to apply KDE for discriminant analysis on high-dimensional datasets.

This new hybrid approach can be easily applied to LDA and QDA by modifying the proposed ICOMP expression and the genetic algorithm slightly. The concept of LDA and QDA are easily comprehensible and these methods are computationally effective. Combining ICOMP, the GA, and LDA or QDA for discriminant analysis will require less computational time than our proposed approach although it may decrease classification accuracy. This approach can be another attractive alternative for discriminant analysis.

For this work, we proposed two research questions. The first research question is whether KDEDA is superior to other methods such as LDA, QDA, and *k*-NNDA. Based on our application to four real data sets, we can conclude that KDEDA performed better than LDA and QDA, and performed as well as *k*-NNDA, with respect to classification error. It is also notable that the nonparametric DA methods including KDEDA and *k*-NNDA performed better than the parametric DA methods such as LDA and QDA. This finding is interesting because some researchers found that nonparametric discriminant functions did not perform as expected, and, in several cases, it performed worse than

parametric discriminant functions (Ferrer & Wang, 1999).

The second research question is whether the new hybrid approach is compatible with the all-possible-subset selection method. Based on our results, it is evident that our approach is compatible with the all-possible-subset selection method for low dimensional data, and it may be compatible with the all-possible-subset selection method for high dimensional data.

In conclusion, our proposed approach has shown excellent performance in predicting group membership and in finding the best model which is as simple as possible. As shown in previously published research in other areas, ICOMP is an attractive model selection criterion for discriminant analysis.

6.2 Future Work

In this dissertation, we only compared our proposed approach with the all-possible-subset selection method. We did not pay attention to other automatic variable selection methods such as stepwise forward and backward methods. These methods are popular and widely used among many researchers. Comparing the effectiveness between our proposed approach and other automatic variable selection methods may be a possible future research topic to study. Second, among several forms of complexity, $C_{1F}\left(\hat{\Sigma}\left(\hat{\theta}\right)\right)$ is utilized for this dissertation. $C_1\left(\hat{\Sigma}\left(\hat{\theta}\right)\right)$ is also an appropriate form of complexity for KDEDA, because both forms are invariant under an orthogonal transformation. Comparing the prediction accuracy of models selected using these two forms of

complexity may help identify models with reduced classification error rates. Third, we found that KDEDA and *k*-NNDA performed better than the parametric DA methods such as LDA and QDA, but there was no significant difference between KDEDA and *k*-NNDA. Some previous research suggested that *k*-NNDA performed worse than LDA and QDA (Ferrer & Wang, 1999). We need to analyze more data sets to compare prediction accuracy between KDEDA and *k*-NNDA.

Bibliography

Aeberhard S, C. D., & O, d. V. (1992). Comparison of classifiers in high dimensional settings. *Technical report no 92-02. Dept. of computer science and Dept. of mathematics and statistics, James Cook University of North Queensland , 92-02*.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csaki (Ed.), *Second international symposium on information theory* (pp. 267-281). Budapest: Academiai Kiado.

Bao, X. (2004). *Computational subset model selection algorithms and applications.* Knoxville.

Bensmail, H., & Bozdogan, H. (2002). Regularized kernel discriminant analysis with optimally scaled data. *Measurement and multivariate analysis* , pp. 133-144.

Blahut, R. E. (1987). *Principles and practice of information theory.* Reading, Massachusetts: Addison-Wesley.

Bowman, A., & Azzalini, A. (1997). *Applied smoothing techniques for data analysis.* London: Oxford university press.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *44*, 62-91.

Bozdogan, H. (1988). ICOMP : A new model-selection criterion. . *In classification and related methods of data analysis* , 599-608.

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in statistics, Part A - Theory and Methods , 19*, 221-278.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new information measure of complexity. In H. Bozdogan, *Multivariate statistical modeling* (pp. 69-113). Dordrecht, the Netherlands: Kluwer Academic Publishers

Bozdogan, H. (2000). Akaike's Information Criterion and Recent Develpments in Information Complexity. *Journal of Mathematical Psychology , 44*, 62-91.

Bozdogan, H. (2004). Intelligent statistical data mining with information complexity and genetic algorithms. In H. Bozdogan, *Statistical data mining and knowledge discovery* (pp. 15-56). New York: Chapman & Hall/CRS.

Bozdogan, H. (2007). Data adaptive kernel discriminant analysis with information complexity. *Tutorial lecture presented at the 6-th scientfic meeting of CLADAG.* Macerata, Italy.

Bozdogan, H. (2009). *Information Complexity in Multivariate Learning and High Dimensional Data Mining. Working book.*

Duong, T., & Hazelton, M. (2003). Plug-in bandwidth slectors for bivariate kernel density estimation. *Journal of nonparametric statistics , 15*, 17-30.

Ferrer, A. J., & Wang, L. (1999). *Comparing the classification accuracy among nonparametric, parametric discriminant analysis and logistic regression methods.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada. (ERIC document reproduction service No. ED 432 591).

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics , 7*, 179-188.

Forster, M. (2000). Key concepts in model selection : performance and generalizability. *Journal of mathematical psychology , 44*, 205-231.

Ghosh, A. K., & Bandyopadhyay, S. (2006). Adaptive smoothing in kernel discriminant analysis. *Nonparametric statistics , 18* (2), 181.

Glen, J. (2001). Classification accuracy in discriminant analysis: a mixed integer programming approach. *Journal of the operational research society , 52*, 328-339.

Goldberg, D. E. (1989). *Genetic algorithms in research, optimization, and machine learning.* Mass.: Addison-Wesley Pub. Co.

Huberty, C. (1994). *Applied discriminant analysis.* New York: Wiley.

Huberty, C. J., & Olejnik, S. (2006). *Applied manova and discriminant analysis.* New York: Wiley.

Hurvich, C. M., & Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika , 76*, 297-307.

Kullback, S., & Leiber, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics , 22*, 79-86.

Kwon, Y. (2003). *Bayesian analysis of threshold autoregressive models (Doctoral dissertation, University of Tennessee, 2003).*

Lin, M., Huang, S., & Chang, Y. (2004). Kernel-based discriminant techniques for educational placement. *Journal of Educational and Behavioral Statistics , 29* (2), 219-240.

Liu, M. (2006). *Multivariate nonnormal regression models, information complexity, and genetic algorithms : a three way hybrid for intelligent data mining.* Knoxivlle.

Liu, Z. (2002). *Intelligent data mining using kernel functions and information criteria (Doctoral dissertation, University of Tennessee, 2002).*

Qiu, X., & Wu, L. (2006). Nearest neighbor discriminant analysis. *International journal of Pattern Recognition* , 20(2), 1245-1259.

Rissanen, J. (1976). Minmax entropy estimation of models for vector process. In R. K. Mehra, & D. G. Lainiotis, *System identification* (pp. 97-119). New York: Academic Press.

Sain, S., Baggerly, K., & Scott, D. (1994). Cross-validation of multivariate densities. *Journal of the american statistical association , 89* (427), 807-817.

Schneider, J. J., & Kirkpatrick, S. (2006). *Stochastic optimization.* New York: Springer.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell systems technology journal , 27*, 379-423.

van Emden, M. H. (1971). *An analysis of complexity* (Vol. 35). Amsterdam: Mathematisch Centrum.

Williams, K. R. (2005). Application of genetic algorithms to a variety of problems in physics and astronomy.(Master's thesis, University of Tennessee, 2005).

Zhang, X., King, M. L., & Hydman, R. J. (2004). Bandwidth selection for multivariate kernel density estimation using MCMC.

Vita

Dong-Ho Park received his bachelor's degree at the Korea Military Academy, South Korea in 1994 and then went to the United States to attend the Naval Postgraduate School in Monterey, CA for his Master of Science and he will finish his doctorate work at the University of Tennessee May 2009.