Statistical Learning Based on High Dimensional Data

M. Stat. Dissertation

by
Shirshendu Chatterjee
(Roll No. MB0414)

Under the supervision of
Prof. Probal Chaudhuri
&
Prof. Debasis Sengupta

Submitted on July 28, 2006

Indian Statistical Institute
203 Barrackpore trunk Road
Kolkata 700108

Contents

Chapter 1

Introduction

1.1 The Classification Problem

The goal of classification problems is to assign objects to one of several (K) clearly identified classes based on a set of measurements $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$. This can be viewed as a statistical decision problems in which we need to partition the sample-space $\mathbb{R}^p$ into K disjoint parts corresponding to different classes. If $f_k(\mathbf{X})$ denotes the class density for the k^{th} class for $k = 1, 2, \ldots, K$, and if we consider 0-1 loss (i.e. no loss for correct classification and one unit loss for each misclassification), the optimum partition (w.r.t. Bayes risk) allocates $\mathbf{X}$ to the i^{th} class if the corresponding posterior probability is maximum, or equivalently $\pi_i f_i(\mathbf{X})$ is maximum, i.e., if

$$\pi_i f_i(\mathbf{X}) = \max_{1 \le k \le K} \pi_k f_k(\mathbf{X}), \tag{1.1.1}$$

where π_k is the prior probability of the k^{th} class.

If we consider normal model, (1.1.1) boils down to minimizing the discriminant score

$$d_k(\mathbf{X}) = (\mathbf{X} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{X} - \boldsymbol{\mu}_k) + \ln |\boldsymbol{\Sigma}_k| - 2 \ln \pi_k \tag{1.1.2}$$

w.r.t. k. But in practice $f_k(.)$'s are seldom known. So we need to estimate the $f_k(.)$'s (in the case of normal model the $\boldsymbol{\mu}_k$s and the $\boldsymbol{\Sigma}_k$s) based on a set of training sample (with N_k items from the k^{th} class) and then construct a classification rule using the estimated densities. Let the training samples be $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})^T$ (j^{th} sample from i^{th} class); $1 \le i \le K, 1 \le j \le N_i$. In QDA, the population parameters are generally estimated by the

corresponding classical unbiased estimators

$$\hat{\boldsymbol{\mu}}_k = \overline{\mathbf{X}}_k = \frac{1}{N_k}\sum_{j=1}^{N_k}\mathbf{X}_{kj} \quad \text{and} \tag{1.1.3}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{N_k - 1}\sum_{j=1}^{N_k}(\mathbf{X}_{kj} - \overline{\mathbf{X}}_k)(\mathbf{X}_{kj} - \overline{\mathbf{X}}_k)^T, \tag{1.1.4}$$

and the k^{th} discriminant scores of a future observation $\mathbf{X}$are estimated by plugging in these estimates of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ in the expression 1.1.2, i.e.

$$\hat{d}_k(\mathbf{X}) = (\mathbf{X} - \hat{\boldsymbol{\mu}}_k)^T\hat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{X} - \hat{\boldsymbol{\mu}}_k) + \ln|\hat{\boldsymbol{\Sigma}}_k| - 2\ln\pi_k. \tag{1.1.5}$$

In LDA, the same thing is done under the assumption that $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}\forall k$ and the common covariance matrix is estimated by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N - K}\sum_{k=1}^{K}\sum_{j=1}^{N_k}(\mathbf{X}_{kj} - \overline{\mathbf{X}}_k)(\mathbf{X}_{kj} - \overline{\mathbf{X}}_k)^T. \tag{1.1.6}$$

So, in this case $\hat{d}_k(\mathbf{X})$ is obtained by replacing $\hat{\boldsymbol{\Sigma}}_k$ in (1.1.5) by $\hat{\boldsymbol{\Sigma}}$.

1.2 The Clustering Problem

The goal of clustering problems is to partition a set of objects in homogeneous groups based on a set of measurements $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p)^T$. Unlike the classification problem, the groups or classes are not identified beforehand, i.e. we need to identify the group structures present in the given set of observations. Clustering may have to be done with or without the help of training data. In the first case, from the training samples we can have the knowledge of homogeneous groups, and also, some idea about the between-group and within-group variations. On the basis of that, we need to discover the homogeneous groups in a different collection of objects, which may be called the test set. Possibly there may be many more homogeneous clusters in the test set, which are not present in the training sample, and the clustering procedure should be able to cope with that. In the absence of training data, we don't have any prior knowledge about the homogeneous groups.

1.3 Problems Arising from High Dimensionality of Data

Both classification and clustering encounter some serious problems, in the case of high dimensional data.

1.3.1 Classification Problems

For classification problems, QDA and LDA perform well if good estimates of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ ($\boldsymbol{\Sigma}$ for LDA) are available. But if the dimension p of the measurement vector is quite large compared to the class sample sizes N_ks, these estimates in (1.1.4) remain no longer reliable. Many of the parameters of the covariance matrices may become unidentifiable. Even if all of them are estimable, the estimates become highly unstable. Thus, we can estimate the discriminant scores only with a very high variance, which leads to high misclassification rates.
This problems may be better understood, if spectral decomposition of the covariance matrices are considered. Let

$$\boldsymbol{\Sigma}_k = \mathbf{V}_k \boldsymbol{\Lambda}_k \mathbf{V}_k^T = \sum_{i=1}^{p} \lambda_{i,k} \mathbf{v}_{i,k} \mathbf{v}_{i,k}^T \tag{1.3.1}$$

be the spectral decomposition of $\boldsymbol{\Sigma}_k$ where $\lambda_{i,k}, 1 \leq i \leq p$, are the eigenvalues of $\boldsymbol{\Sigma}_k$ (in descending order) with corresponding eigenvectors $\mathbf{v}_{i,k}$. So,

$$\boldsymbol{\Sigma}_k^{-1} = \sum_{i=1}^{p} \frac{1}{\lambda_{i,k}} \mathbf{v}_{i,k} \mathbf{v}_{i,k}^T. \tag{1.3.2}$$

Hence,

$$d_k(\mathbf{X}) = \sum_{i=1}^{p} \frac{[\mathbf{v}_{i,k}^T (\mathbf{X} - \boldsymbol{\mu}_k)]^2}{\lambda_{i,k}}. \tag{1.3.3}$$

Clearly, $d_k(\mathbf{X})$ is heavily weighted by the small eigenvalues of $\boldsymbol{\Sigma}_k$. When $d_k(\mathbf{X})$ is estimated as above, then $\mathbf{v}_{i,k}$ and $\lambda_{i,k}$ are estimated by the corresponding eigenvalues and eigenvectors of the $\hat{\boldsymbol{\Sigma}}_k$, i.e.,

$$\hat{d}_k(\mathbf{X}) = \sum_{i=1}^{p} \frac{[\hat{\mathbf{v}}_{i,k}^T (\mathbf{X} - \hat{\boldsymbol{\mu}}_k)]^2}{\hat{\gamma}_{i,k}}. \tag{1.3.4}$$

Now, when N_k is comparable to p, the small eigenvalues of $\mathbf{\Sigma}_k$ are underestimated and the corresponding estimates become highly unstable, as mentioned in Friedman [8]. In the extreme case, when $N_k < p$ the last $p - N_k$ of the $\hat{\gamma}_{i,k}$'s are 0, so QDA is no longer useful. As a consequence, the direction associated with a small $\lambda_{i,k}$ gets over-importance, which increases the variance of $\hat{d}_k(\mathbf{X})$. Thus the misclassification rates get enhanced.

1.3.2 Clustering Problems

If the dimension p of the measurement vector is substantially large compared to the number of the objects N to be partitioned, clustering with or without training data may become difficult. In most of the high dimensional clustering problems, the true cluster structure remains confined to much lower dimensional subspaces. Inclusion of too many noisy variables, which are less relevant in clustering, may hinder the recovery of the actual homogeneous groups present in the clusters. For example, if hierarchical clustering algorithm is used in case of high dimensional data, different homogeneous groups may become unidentifiable, unless the groups are very far away from each other in terms of their Mahalanobis distances. A simulation study regarding this fact is discussed in Chapter 5. In the case of model based clustering also, if the dimension is very large, the estimate of within-group variation and between-group variation which we get, is highly unstable and unreliable. Naturally, partitioning the objects based on these estimates leads to high misclassification rates. The details are discussed in Chapter 6.

1.4 Available Approaches for Classifying High Dimensional Data

To cope with the difficulties in using QDA for those problems, where the number of observations is either marginally more or less than the number of unknown parameters (poorly-posed or ill-posed problems, respectively), there are some techniques available in the literature. Broadly two kinds of approaches can be adopted for these problems. One approach tries to obtain reliable and more stable estimates of the parameters without changing the model, and the other approach imposes restrictions on the model to reduce the number of unknown parameters. In some situations, the second approach is more useful compared to the first one.

Bickel and Levina [4], have shown that, in the case of the classification problems, in which the dimension p is quite large compared to the sample sizes, rules which use evidently incorrect assumption that the variables are independent, often perform better than rules which try to capture dependence structure in the construction of the classifier (detail clarification in section 2.4). This gives a motivation to go for restrictive models. In the following subsections, some methods involving both the approaches are mentioned.

1.4.1 More Accurate Estimation of Eigenvalue

One possible way to deal with the problem due to high dimensionality is to estimate the eigenvalues of the covariance matrices more accurately, reducing the bias involved in the classical estimates. James and Stein [13], Effron and Morris [6], Lin and Perlman [14], and Dey and Srinivasan [5] tried to get estimates of the eigenvalues by minimizing certain loss functions. But none of these loss functions are constructed to minimize the misclassification risk of a classification problems. Also, $\hat{\mathbf{\Sigma}}_k$ is required to be nonsingular in almost all of these approaches.

1.4.2 Variable Selection and Dimension Reduction

Another possible way to cope with high dimensional data is variable selection, as described in Schaafsma [16]. In this approach, one tries to choose judiciously a smaller subset of variables, which have stronger discriminating power, discarding the less relevant ones, which cannot distinguish the populations well. Generally, the optimal subset of variables for a particular subset size is obtained by minimizing the misclassification rates over all possible subsets of that size. Then the optimal subsets of different sizes are compared on the basis of their misclassification rates, to determine the most optimal one. If the variables can be ordered according to their importance in classification, the optimal subsets can be found more easily because in that case the first few variables (of the ordered set) will be optimal subset of the corresponding subset size. Alternatively, dimension reduction techniques (using linear combination of the variables or other similar methods) may also be considered. In this context, principal components are often used to reduce the dimension. Since a natural ordering exists among the principal components, it is required to find out only the number of components to be used in a particular problems. Generally, in the situations where the variables are ordered, the misclassification rate decreases with the increase of size of the ordered subset. So, one tries to get the smallest subset such that the

corresponding misclassification probability is not substantially different from that of the full set of variables. This choice is subjective .

1.4.3 Regularization Techniques

Regularization techniques are very useful and popular in the situation of high dimensional classification problems. A regularization technique typically consists of shrinking an estimated covariance matrix towards a matrix of specified form.[12] In the following subsections some important regularization techniques are discussed.

Regularized Discriminant Analysis (RDA)

Friedman [8] proposed some regularization method called Regularized Discriminant Analysis (RDA). RDA tries to improve the classical estimates of the class covariance matrices by reducing their variance at the cost of biasing them away from the sample based estimates towards a more plausible set of values. The increase in bias mainly depends on how close the actual population covariance matrices and the plausible set of values are. If the plausible set closely approximates the actual parameters, substantial reduction in variance can be achieved at the expense of a small increase in bias. This bias-variance trade-off is controlled by two regularization parameters γ and γ $(0 \leq \gamma, \gamma \leq 1)$. The complexity parameter γ controls the regularization of $\mathbf{\Sigma}_k$ towards the pooled covariance matrix $\mathbf{\Sigma}$.

$$\mathbf{\Sigma}_k(\gamma) = \frac{(1-\gamma)\mathbf{S}_k + \gamma\mathbf{S}}{(1-\gamma)N_k + \gamma N}, \tag{1.4.1}$$

where $\mathbf{S}_k = N_k\hat{\mathbf{\Sigma}}_k$, $N = \sum_{i=1}^{K} N_k$ and $\mathbf{S} = \sum_{i=1}^{K} \mathbf{S}_k$. The other parameter γ is used to further regularize the class covariance matrices and controls their shrinkage towards a suitable multiple of the identity matrix.

$$\hat{\mathbf{\Sigma}}_k(\gamma, \gamma) = (1-\gamma)\hat{\mathbf{\Sigma}}_k(\gamma) + \gamma \left[\frac{trace\{\hat{\mathbf{\Sigma}}_k(\gamma)\}}{p}\right] \mathbf{I}. \tag{1.4.2}$$

The shrinkage towards a multiple of the identity matrix helps to counteract the bias involved in estimating the eigenvalues. Thus RDA suggests an intermediate classifier among 4 classification rules corresponding to QDA $(\gamma = 0, \gamma = 0)$, LDA $(\gamma = 1, \gamma = 0)$, classifiers based on a model with $\mathbf{\Sigma}_k = \sigma_k^2\mathbf{I}$ $(\gamma = 0, \gamma = 1)$ and a model with $\mathbf{\Sigma}_k = \sigma^2\mathbf{I}$ $(\gamma = 1, \gamma = 1)$, respectively. Sample based estimates of these parameters are obtained by minimizing the misclassification risk which is obtained using cross-validation method.

Eigenvalue Decomposition Discriminant Analysis (EDDA)

In another regularization approach, called Eigenvalue Decomposition Discriminant Analysis (EDDA) Bensmail and Celeux [2] considered reparametrization of the covariance matrices with respect to eigenvalue decomposition. If $\mathbf{\Lambda}_k$ is written as

$$\mathbf{\Lambda}_k = \alpha_k \mathbf{A}_k, \tag{1.4.3}$$

where $\mathbf{\Lambda}_k$ is as in (1.3.1) and $\alpha_k = (\prod_{i=1}^{p} \lambda_{i,k})^{\frac{1}{p}}$, then

$$\mathbf{\Sigma}_k = \alpha_k \mathbf{V}_k \mathbf{A}_k \mathbf{V}_k^T. \tag{1.4.4}$$

Here $\mathbf{A}_k$ is a diagonal matrix consisting of standardized eigenvalues of $\mathbf{\Sigma}_k$, and α_k, $\mathbf{V}_k$ and $\mathbf{A}_k$ represent respectively the volume, the orientation and the shape of the k^{th} population density. For various constraints on these parameters, e.g., varying some but not all the parameters across the population, different discrimination models were obtained (this includes LDA and QDA as special cases). After estimating the parameters by the corresponding m.l.e., the model which leads to the least future misclassification risk, was chosen to form the classification rule.

High Dimensional Discriminant Analysis (HDDA)

There is another approach called High Dimensional Discriminant Analysis (HDDA) [Bouveyron, Girard, Schmid, 2005], which is based on the assumption that high dimensional data lives in different low dimensional subspaces. In this approach, the dimension of different classes are reduced independently and then the class covariance matrices are regularized to adopt the Gaussian framework. Here regularization is based on the assumption that the classes have spherical eigen-spaces.

Some Theoretical Results Favoring Regularization

Bickel and Levina [4] have shown theoretically that in the case of high dimensional classification problems with two Gaussian populations having same covariance matrix and different mean vectors, based on a sample of size n from each population, the rules which assume the variables to be independent, perform much better than those based on covariance matrix that tries to capture the correlation among variables, under certain broad conditions namely $\frac{p}{n} \to \infty$ and $\frac{log(p)}{n} \to 0$ as $n \to \infty$ and eigenvalues of $\mathbf{\Sigma}$ do not converge to 0 or ∞ as $p \to \infty$. This includes the case when $p = \mathrm{O}\big(n^{\delta}\big)\, \forall \delta > 1$. Since in the asymptotic analysis p also tends to ∞ along with n, the mean

vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ can be made elements of the space l_2 by adding 0s at the end. In the standard LDA, $\boldsymbol{\Sigma}^{-1}$ is replaced by $\boldsymbol{\Sigma}^{+}$ (the Moore-Penrose G-inverse of $\boldsymbol{\Sigma}$) whenever $n > p$. The parameter space, which was considered is

$$\Gamma(c, k_1, k_2, B) = \{(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) : \Delta \geq c^2; k_1 \leq \lambda_{min}(\boldsymbol{\Sigma}) \leq \lambda_{max}(\boldsymbol{\Sigma}) \leq k_2; \boldsymbol{\mu} \in B\}, \tag{1.4.5}$$

where $c,\ k_1$ and k_2 are some positive numbers, Δ is the Mahalanobis distance between the two populations, and B is a compact subset of the space l_2 of the form

$$B = B_{a,d} = \left\{ \boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots) : \sum_{j=1}^{\infty} a_j \boldsymbol{\mu}_j^2 \leq d \right\}. \tag{1.4.6}$$

The form of the parameters ensures that none of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ is ill-conditioned and c represents the difficulty level of the problems. Clearly, the Bayes Risk of this problems is $\overline{\Phi}(c/2)$, where $\overline{\Phi}(.) = 1 - \Phi(.)$ and $\Phi(.)$ is the cdf of standard normal distribution.
Let $\mathbf{D}$ be the diagonal matrix with the same diagonal elements as $\boldsymbol{\Sigma}$. If $\mathbf{X}$ is a future observation from $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, and $M_{LDA}(\theta)$ and $M_I(\theta)$ represent the (unconditional) probabilities of misclassifying $\mathbf{X}$ corresponding to the standard LDA and the independent rule respectively, i.e.,

$$M_{LDA}(\theta) = P_\theta \left[\left(\mathbf{X} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2} \right)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) < 0 \right] \tag{1.4.7}$$

$$M_I(\theta) = P_\theta \left[\left(\mathbf{X} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2} \right)^T \hat{\mathbf{D}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) < 0 \right] \tag{1.4.8}$$

and $M_{LDA,\Gamma},\ M_{I,\Gamma}$ represent corresponding worst performance, i.e.,

$$M_{LDA,\Gamma} = \max_{\Gamma} M_{LDA}(\theta) \tag{1.4.9}$$

$$M_{I,\Gamma} = \max_{\Gamma} M_I(\theta), \tag{1.4.10}$$

then it has been proved that

$$\frac{p}{n} \to \infty \Rightarrow M_{LDA,\Gamma} \to \frac{1}{2}, \tag{1.4.11}$$

$$\frac{log(p)}{n} \to 0 \Rightarrow \limsup_{n \to \infty} M_{I,\Gamma} = \overline{\Phi} \left(\frac{\sqrt{k_0}}{1 + k_0} c \right), \tag{1.4.12}$$

where

$$k_0 = \max_{\Gamma} \frac{\lambda_{max}(\mathbf{\Sigma}_0)}{\lambda_{min}(\mathbf{\Sigma}_0)} \tag{1.4.13}$$

with $\mathbf{\Sigma}_0 = \mathbf{D}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{D}^{-\frac{1}{2}}$ as the correlation matrix corresponding to $\mathbf{\Sigma}$. This result suggests that if the eigenvalues of $\mathbf{\Sigma}$ do not converge to 0 nor do they diverge to ∞ as $p \to \infty$, i.e., k_0 is finite; then the worst performance of the "independent rule" is a decreasing function of the Mahalanobis distance, while the standard LDA is no better than random guessing irrespective of the value of the Mahalanobis distance between the two populations.
The results also suggest that, under the assumptions on n and p, all variables can be used for classification ignoring the correlations. This gives a motivation to go for shrinkage towards a diagonal matrix because in that approach, the "independent rule" is included as a special case for a specific value of the shrinkage parameter. This approach will be discussed in the next chapter.

1.5 Available Approaches for Clustering High Dimensional Data

In the context of high dimensional clustering problems, a common variable selection method involves fitting univariate models to each component and choosing those which satisfy some threshold criteria. This was proposed by McLachlan et al. [15]. But this method does not take into account the joint effect of the variables. So, the variables, which help in clustering in presence of others, but individually are not potentially important, are ignored. Another standard dimension reduction technique is to use leading principal components and then find the clusters using any standard procedure, as described in [11]. In these two approaches, the variable selection and clustering are done separately. There are some methods in which these two tasks are done simultaneously. For example, forward selection methods can be used in the case of hierarchical algorithms. This was proposed by Fowlkes et al. [7]. In this approach, variables are included based on the information about between-cluster and total sum of squares. The significance of the variables was judged on the basis of graphical information. In another approach, proposed by Mahlet et al. [18], the clustering problems was formulated in terms of a multivariate mixture model with an unknown number of components. Then MCMC techniques were used to infer about the selection of discriminating variables, estimates for the number of clusters and the sample allocations.

Chapter 2

Preliminary Study of Some Classification Rules

Initially, two variable selection methods and some regularization techniques were considered. These are discussed in the next two sections.

2.1 Variable Selection Approach

In both the variable selection methods, the performance of a subset of variables (for classification) was measured by means of two estimates of the ratio of mean squared neighbor distance and mean squared non-neighbor distance. Here, the mean neighbor distance is the expected Euclidian distance within the same population, and the mean non-neighbor distance is the expected Euclidean distance between two populations. These two estimates are (1) average of ratios (R_1) and (2) ratio of averages (R_2) as given below.

$$R_1 = \frac{1}{N}\sum_{i=1}^{K}\sum_{j=1}^{N_i} R_{1,i,j}, \tag{2.1.1}$$

where,

$$R_{1,i,j} = \frac{\sum_{j'=1 j'\neq j}^{N_i} \left\|X_{i,j} - X_{i,j'}\right\|^2}{\sum_{i'=1,i'\neq i}^{K}\sum_{j'=1}^{N_{i'}} \left\|X_{i,j} - X_{i',j'}\right\|^2}, \tag{2.1.2}$$

and

$$R_2 = \frac{\left(\sum_i N_i(N_i-1)\right)^{-1}\sum_{i,j,j';j\neq j'} \left\|X_{i,j} - X_{i,j'}\right\|^2}{\left(\sum_i N_i(N-N_i)\right)^{-1}\sum_{i,i',j,j';i\neq i'} \left\|X_{i,j} - X_{i',j'}\right\|^2}. \tag{2.1.3}$$

In the first variable selection method, optimal subsets of variables of different sizes (starting from 1 and going up to p) were obtained using the above criterion and using forward selection method. Different subsets were compared on the basis of the two ratio estimates.
In the second method of variable selection, first the variables were sorted in ascending order of ratio estimates. Then the subset of variables consisting of first d variables were considered, $1 \leq d \leq p$, as the optimal d-dimensional subset. Then, using LDA, corresponding misclassification rates were compared.

2.2 Regularization Approach

In this section, we discuss some regularization techniques. For the time being, we assume that $K = 2$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$.

2.2.1 Shrinkage Towards Multiple of Identity Matrix

Under the above assumption, a special case of RDA was considered. It uses shrinkage of the common covariance matrix towards a multiple of the identity matrix, where the multiple is $p^{-1}\text{trace}(\hat{\mathbf{\Sigma}})$, which is also the MLE of σ^2 under the model $\mathbf{\Sigma}_k = \sigma^2 I \ \forall\, k$, where $\hat{\mathbf{\Sigma}}$ is given by (1.1.6). This shrinkage tries to pull all the eigenvalues towards their average, and thereby try to cope with the underestimation of the small eigenvalues of $\mathbf{\Sigma}$.

2.2.2 Shrinkage Towards Diagonal Matrix

In our work, along with the above method, we have also considered two more shrinkage methods. In one of the methods, as motivated at the end of section 1.4.3, we try to regularize the common covariance matrix by shrinking it towards a suitable diagonal matrix. The amount of shrinkage is controlled by a parameter γ. The diagonal matrix is taken to be the MLE of $\mathbf{\Sigma}$ under the assumption that $\mathbf{\Sigma}$ is diagonal. This MLE is given by $\hat{\mathbf{D}}$ =DIAG($\hat{\mathbf{\Sigma}}$), i.e., a diagonal matrix having the same diagonal entries as of $\hat{\mathbf{\Sigma}}$. So, the regularization is given by

$$\hat{\mathbf{\Sigma}}_D(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\hat{\mathbf{D}} \tag{2.2.1}$$

This method actually keeps the variance estimates of the variables unchanged and reduces the covariance estimates uniformly. This method also tries to resolve the eigenvalue distortion problem as shown below. For the smallest eigenvalue of $\hat{\mathbf{\Sigma}}_D(\gamma)$ we have

$$\lambda_{min}(\hat{\mathbf{\Sigma}}_D(\gamma)) = \inf_{\|\mathbf{x}\|=1} \mathbf{x}^T\hat{\mathbf{\Sigma}}_D(\gamma)\mathbf{x} = \inf_{\|\mathbf{x}\|=1} \left[(1-\gamma)\mathbf{x}^T\hat{\mathbf{\Sigma}}\mathbf{x} + \gamma\mathbf{x}^T\hat{\mathbf{D}}\mathbf{x}\right]$$
$$\geq (1-\gamma) \inf_{\|\mathbf{x}\|=1} \mathbf{x}^T\hat{\mathbf{\Sigma}}\mathbf{x} + \gamma \inf_{\|\mathbf{x}\|=1} \mathbf{x}^t\hat{\mathbf{D}}\mathbf{x} \geq \inf_{\|\mathbf{x}\|=1} \mathbf{x}^T\hat{\mathbf{\Sigma}}\mathbf{x} = \lambda_{min}(\hat{\mathbf{\Sigma}}). \quad (2.2.2)$$

Similarly, it reduces the maximum eigenvalue. If the variables are indeed independent, the bias introduced due to regularization will be quite small, while great reduction in variance of the discriminant scores can be achieved.

2.2.3 Shrinkage Towards Intra-class Correlation Matrix

In the other method, shrinkage towards a suitable intraclass correlation matrix (which is the MLE of $\mathbf{\Sigma}$ under the assumption that all diagonal entries of $\mathbf{\Sigma}$ are σ^2 and all off-diagonal entries are $\sigma^2\rho$). This assumption on $\mathbf{\Sigma}$ to have the intraclass correlation structure is a good compromise between the assumption of arbitrary covariance structure and ignoring the correlations. This assumption tries to capture the average correlation among the variables instead of estimating all of them. Under this assumption, the MLE of $\mathbf{\Sigma}$ is given by

$$\hat{\mathbf{\Sigma}}_C = \hat{\sigma^2}(1-\hat{\rho})\mathbf{I}_p + \hat{\sigma^2}\hat{\rho}\mathbf{J}_p, \quad (2.2.3)$$

where

$$\hat{\sigma^2} = \frac{trace(\hat{\mathbf{\Sigma}})}{p}, \quad (2.2.4)$$

$$\hat{\sigma^2}\hat{\rho} = \frac{\sum_{i,j=1;i\neq j}^{p} \hat{\mathbf{\Sigma}}(i,j)}{p(p-1)}. \quad (2.2.5)$$

So the regularization is given by

$$\hat{\mathbf{\Sigma}}_C(\gamma) = (1-\gamma)\hat{\mathbf{\Sigma}} + \gamma\hat{\mathbf{\Sigma}}_C. \quad (2.2.6)$$

This regularization also improves the eigenvalues like the other one, because the distinct eigenvalues of $\hat{\mathbf{\Sigma}}_C$ are only $\hat{\sigma}^2(1-\hat{\rho})$ and $[\hat{\sigma}^2 + (p-1)\hat{\sigma}^2\hat{\rho}]$. The second eigenvalue is nothing but $\underline{\mathbf{1}}^T\hat{\mathbf{\Sigma}}\underline{\mathbf{1}}/\underline{\mathbf{1}}^T\underline{\mathbf{1}}$, and

$$\frac{\underline{\mathbf{1}}^T\hat{\mathbf{\Sigma}}\underline{\mathbf{1}}}{\underline{\mathbf{1}}^T\underline{\mathbf{1}}} \geq \min_{\mathbf{l}\neq 0} \frac{\mathbf{l}^T\hat{\mathbf{\Sigma}}\mathbf{l}}{\mathbf{l}^T\mathbf{l}} = \lambda_{min}(\hat{\mathbf{\Sigma}}).$$

The other eigenvalue is

$$
\begin{aligned}
&\hat{\sigma^2}(1-\hat{\rho}) \\
&= \frac{\text{trace}(\hat{\mathbf{\Sigma}})}{p} - \frac{\underline{\mathbf{1}}^T\hat{\mathbf{\Sigma}}\underline{\mathbf{1}} - \text{trace}(\hat{\mathbf{\Sigma}})}{p(p-1)} \\
&= \frac{\sum_{i=1}^{p}\mathbf{e}_i^T\hat{\mathbf{\Sigma}}\mathbf{e}_i}{p} - \frac{\underline{\mathbf{1}}^T\hat{\mathbf{\Sigma}}\underline{\mathbf{1}} - \sum_{i=1}^{p}\mathbf{e}_i^T\hat{\mathbf{\Sigma}}\mathbf{e}_i}{p(p-1)} \text{ [where, the } \mathbf{e}_i\text{'s are the unit vectors of } \mathbb{R}^P] \\
&= \frac{1}{p-1}\left[\sum_{i=1}^{p}\mathbf{e}_i^T\hat{\mathbf{\Sigma}}\mathbf{e}_i - \frac{1}{p}\underline{\mathbf{1}}^T\hat{\mathbf{\Sigma}}\underline{\mathbf{1}}\right] \\
&= \frac{1}{p-1}\sum_{i=1}^{p}\left[\text{trace}(\hat{\mathbf{\Sigma}}\mathbf{e}_i\mathbf{e}_i^T)\right] - \text{trace}(\hat{\mathbf{\Sigma}}\frac{\underline{\mathbf{1}}\underline{\mathbf{1}}^T}{p}) \\
&= \frac{1}{p-1}\text{trace}\left[\hat{\mathbf{\Sigma}}\left(\sum_{i=1}^{p}\mathbf{e}_i\mathbf{e}_i^T - \frac{\underline{\mathbf{1}}\underline{\mathbf{1}}^T}{p}\right)\right] \\
&= \frac{1}{p-1}\text{trace}\left[\hat{\mathbf{\Sigma}}\left(\mathbf{I}_p - \frac{\mathbf{J}_p}{p}\right)\right] \\
&= \frac{1}{p-1}\text{trace}\left[\hat{\mathbf{\Sigma}}\left(\sum_{i=1}^{p-1}\mathbf{u}_i\mathbf{u}_i^T\right)\right] \\
&\text{(where, the } \mathbf{u}_i\text{'s are the eigenvectors of } \mathbf{I}_p - \frac{\mathbf{J}_p}{p} \text{ corresponding to the eigenvalue 1)} \\
&= \frac{1}{p-1}\sum_{i=1}^{p-1}\text{trace}\left(\mathbf{u}_i^T\hat{\mathbf{\Sigma}}\mathbf{u}_i\right) \\
&\geq \frac{1}{p-1}\sum_{i=1}^{p-1}\lambda_{min}(\hat{\mathbf{\Sigma}}) \\
&= \lambda_{min}(\hat{\mathbf{\Sigma}}). \qquad (2.2.7)
\end{aligned}
$$

So, $\lambda_{min}(\hat{\mathbf{\Sigma}}_C) \geq \lambda_{min}(\hat{\mathbf{\Sigma}})$. Hence, using the same argument as in (28), $\lambda_{min}(\hat{\mathbf{\Sigma}}_C(\gamma)) \geq \lambda_{min}(\hat{\mathbf{\Sigma}})$.
In all these regularization methods, the optimal shrinkage parameter is estimated by minimizing the misclassification rate w.r.t. γ. To do this, the interval [0, 1] is partitioned using s many equally spaced grid points : $0 = t_0, t_1, \ldots, t_s = 1$ (s is decided beforehand). Then the misclassification rate corresponding to each γ ($\gamma = t_0, t_1, \ldots, t_s$) is obtained by using the leave-one-out cross-validation method. The t_i, for which the rate is minimum, is taken as the value of γ.

Smooth curve of misclassification Rates

There is one problem with the above method of estimation of the shrinkage parameter. If the leave-one-out cross-validation method is used to estimate the misclassification probability, the estimate turns out to be same for many γs. So, the optimal γ cannot be determined uniquely. Also, the minimizer depends on the number of grid points s. To overcome this difficulty, several independent random partitions (into two parts) of the training set were used. For each of these partitions, the estimates of misclassification probabilities for all the partition-points were obtained by classifying one part of the training data using the classifiers (corresponding to each grid point) based on the other part. A much smoother curve of the misclassification probability is obtained by averaging the misclassification rates over these random partitions. If the minimizer is still not unique, we go for the largest one.

2.3 Regularization with Aggregation

An alternative approach is to combine the results of all γ's, instead of choosing a particular γ, which minimizes the misclassification probability estimate. If the results obtained at different levels of regularization are combined the performance may improve. A natural choice of aggregation is to consider a suitable weighted average of the posterior probabilities corresponding to various γ. Bagging (see e.g. Breiman [3]), Boosting (see e.g. Schapire et al. [17], Friedman, Hastie and Tibshirani [9]) are some well known aggregation methods, which have been successfully used to combine the results of different classifier. The weighted posterior probability (function of γ) for each class is integrated over γ to obtain a new posterior probability. Then using this posterior, Bayesian classification rule can be constructed.
In our study we have considered two weight functions, similar to those used in Ghosh et al. [10]. However, the simulation study (see Section 2.4.2) shows that the aggregation methods do not improve the misclassification rates.

2.4 Simulation Studies

2.4.1 For Variable Selection

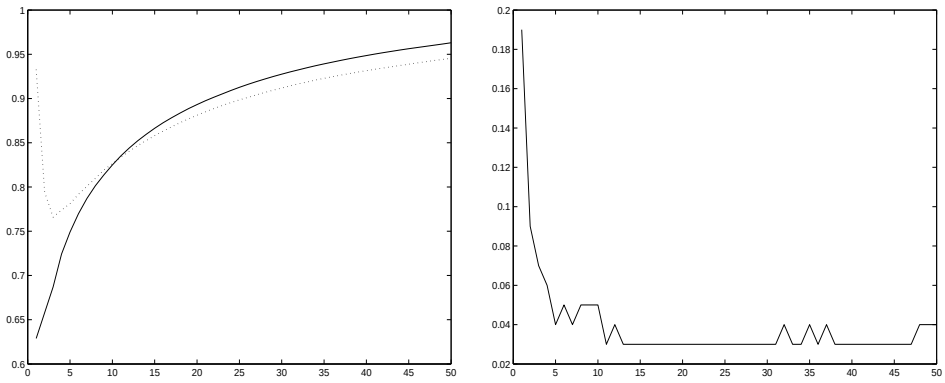
Simulation Plan

In the numerical experiment, two samples, each of size 50 from $N_{100}(\mathbf{0}, \mathbf{I}_{100})$ and $N_{100}(\boldsymbol{\mu}, \mathbf{I}_{100})$ were considered, where the first ten components of $\boldsymbol{\mu}$ are

$(.2, .4, \ldots, 2)$ and all other components are zeros. For the numerical experiment of the second method, the same example was considered.

Results

Both the ratio estimates were observed to increase with the increase of dimension of the optimal subsets (as shown in (2.4.1)). So, this method may not perform well in selection of variables, as the ratio estimates attain minimum for subset size 1.



(a) (b)

Figure 2.4.1: Plots of (a) ratio estimate against dimension and (b) misclassification rate against dimension.

In the second method, the misclassification rates initially decreases with the increase of dimension of the optimal subset and more or less stabilizes after some stage, as shown in Figure 2.4.1(b). In this method also, no clear guideline was obtained for discarding the less relevant variables, namely the variables with the same mean for the two populations.

Discussion

While investigating this particular behavior of the ratio estimates, it was realized that this behavior of R_2 is quite expected. The reason for this behavior can be explained as follows. In this case numerator of R_2 is the average of $N_1 + N_2$ random variables each having $\chi_2(d)$ distribution, if subset of d variables is considered. Similarly, the denominator of R_2 is the average of $2N_1N_2$ random variables each following noncentral $\chi_2(d, \nu^T\nu)$ distribution, where ν is the separation between the mean of the two populations corresponding to the subset of d variables. So, if the numerator and the denominator of R_2 can be approximated by the mean of the corresponding distribution, the

population neighbor vs. non-neighbor will be approximately $2d/(2d+\nu^T\nu)$, where $\nu = (\nu_1, \ldots, \nu_d)$. Now if we include another variable in the subset, the corresponding population ratio becomes $(2d+2)/(2d+2+\nu_1^T\nu_1)$, where $\nu_1 = (\nu_1, \ldots, \nu_d, \nu_{d+1})$. So there will be an improvement after the inclusion, only if

$$\frac{2d}{2d+\nu^T\nu} > \frac{2(d+1)}{2(d+1)+\nu_1^T\nu_1} \Leftrightarrow \nu_{d+1} > \frac{\nu^T\nu}{d}. \tag{2.4.1}$$

This condition will not be satisfied if the variables are arranged in decreasing order of mean-separation between the two populations and the first d variables are considered as the optimal d-subset. In our experiment, after choosing the variables using forward selection method, they become ordered in the previous sense, and hence the increasing behavior of R_2 is observed. Later investigation proved that this is also not very surprising. Because, if LDA is used with known mean and covariance matrix, the actual misclassification probability is given by $\overline{\Phi}(\Delta)$, where

$$\Delta = \mu^T\mathbf{\Sigma}^{-1}\mu \tag{2.4.2}$$

is the Mahalanobis distance between the two populations and μ is the difference between the two populations. In our example, $\Delta = \mu^T\mu$. So, whenever a variable is included in a subset, Δ will increase (unless the included variable has the same mean for the populations) resulting in decrease of misclassification probability, irrespective of whether the included variable is relevant or not. So, this method also does not provide a unique optimal subset of variables. The only possible way to discard the irrelevant variables is to get the value of d after which there is no significant change in the misclassification rates in the above method. This will be considered later.

A Conceptual Issue

Another important fact regarding variable selection is that, if the variables are correlated, the apparently irrelevant variables (i.e. having equal mean for the populations) may sometime improve the misclassification probability considerably. To see this, let us consider a partition

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1^T \\ \mu_2^T \end{bmatrix} \text{ and } \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix} \tag{2.4.3}$$

where $\mu_2 = 0$ and μ_1 is $d \times 1$. Now, if we use only the first d variables for classification, then the Bayes Risk will be $\overline{\Phi}(\Delta_1)$ and if all the variables are

used, then the Bayes Risk will be $\overline{\Phi}(\Delta_2)$, where

$$\Delta_1 = \boldsymbol{\mu}^T \boldsymbol{\Sigma}_{11}^{-1} \mu_1 \tag{2.4.4}$$

$$\Delta_2 = \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{11} \mu_1, \tag{2.4.5}$$

and

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{11} & \boldsymbol{\Sigma}^{12} \\ \boldsymbol{\Sigma}^{21} & \boldsymbol{\Sigma}^{22} \end{bmatrix}. \tag{2.4.6}$$

Clearly, $\boldsymbol{\Sigma}^{11} = (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})^{-1} \geq \boldsymbol{\Sigma}_{11}^{-1}$ (in Lowner order). So, $\Delta_1 \leq \Delta_2$ and hence $\overline{\Phi}(\Delta_1) \geq \overline{\Phi}(\Delta_2)$. This phenomenon is also reflected in the numerical experiment. To see the improvement of misclassification rate, when all the variables are used, two experiments were done. In the first one, $N_2(0, \boldsymbol{\Sigma})$ and $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ were considered as the populations, where $\boldsymbol{\mu}^T = [2, 0]$, $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} = 1$. For various choices of $\boldsymbol{\Sigma}_{12}$, misclassification rates corresponding to the two classifiers (based on only the first variable and both the variables, respectively) were compared (see (2.4.2)). In the other one, two populations were considered from $N_3(0, \boldsymbol{\Sigma})$ and $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution, where $\boldsymbol{\mu}^T = (2, 0, 0)$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix} \tag{2.4.7}$$

Misclassification using all three variables is observed to be considerably less than that of the classifier, which uses only the first variable. The difference is more prominent if ρ is more than .5 (see (2.4.2)). This is a difficult aspect of the variable selection method, because seemingly irrelevant variables can sometimes help to improve misclassification rates. In these situations one should not drop those variables.

For this reason the variable selection method, I have tried some alternatives like Regularization.

2.4.2 For Regularization

Simulation Plan

To compare the performances of three different shrinkage methods namely shrinkage towards multiple of identity, intra-class correlation and diagonal matrix respectively and three different procedures (minimization of the misclassification rates w.r.t the shrinkage parameter, general aggregation and

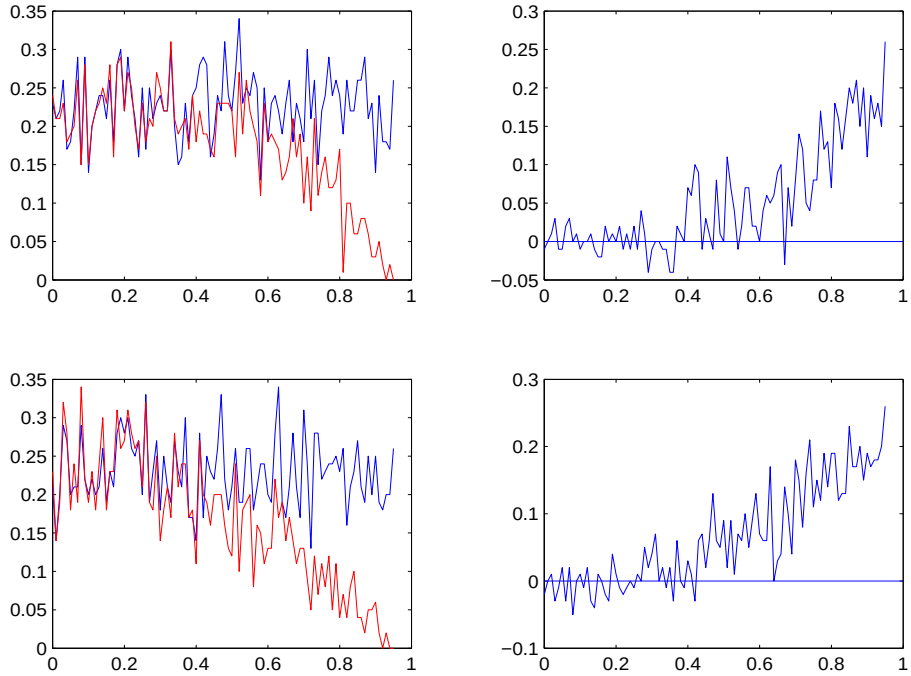


Figure 2.4.2: Decrease in misclassification probability, when all the variables are used instead of the only one which has different means for the two populations

case-dependent aggregation) associated with each of the methods, for various combinations of the parameters of the true underlying populations, the following numerical experiment was conducted. For fixed n (sample size per population), d (dimension of the populations) and Δ (the squared Mahalanobis distance between the two populations, which represents the difficulty level of classification) 6 different covariance matrices were considered. They are

$\mathbf{\Sigma}_1$ = Identity matrix of order d
$\mathbf{\Sigma}_2$ =Intraclass Correlation Matrix with each variance 1 and correlation 0.1
$\mathbf{\Sigma}_3$ = Intraclass Correlation Matrix with each variance 1 and correlation 0.5
$\mathbf{\Sigma}_4$ = Intraclass Correlation Matrix with each variance 1 and correlation 0.9
$\mathbf{\Sigma}_5$ = Diagonal Matrix with variances $1, 2, \ldots, d$
$\mathbf{\Sigma}_6$ = Block Diagonal Matrix with three blocks of sizes $[\frac{d}{3}], [\frac{d}{3}]$ and $d - 2[\frac{d}{3}]$ where the first block has intra-class correlation structure with variance d and covariance $d - 1$, the second block is a diagonal matrixwith variances $1, 2, \ldots, [\frac{d}{3}]$ and the last block is an identity matrix of corresponding order

The mean separation between the two populations was taken as

$$k\left(1, 2, \ldots, \left[\tfrac{d}{3}\right], 0, 0, \ldots, 0\right),$$

for some constant k, which is uniquely determined by Δ and $\boldsymbol{\Sigma}_i, i = 1, 2, \ldots, 6$. For each $\boldsymbol{\Sigma}_i, i = 1, 2, \ldots, 6$, two training sets, each of size n, from $N_d\left(0, \boldsymbol{\Sigma}_i\right)$ and $N_d\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_i\right)$ were considered. The interval $[0, 1]$ is divided into 100 grid points (including 0 and 1) for γ. Smooth curve of misclassification rates corresponding to the three different shrinkage method were obtained averaging over 100 independent random partitions of the training sets and using the method described in section (3.2.4). Misclassification probabilities for 9 classifiers were obtained based on the full training set and using the above smooth curve. Lastly, misclassification rates corresponding to the classifiers were obtained by applying them on an independent test set of sample size 1000. The experiment was repeated for $\Delta = 1, 4, 9, 16$ (for fixed n and d). Then, the whole experiment is repeated for several n and d combinations. For fixed n and Δ, a 3×2 panel was used to plot the misclassification rates, where i^{th} panel (viewed row- wise)corresponds to $\boldsymbol{\Sigma}_i, i = 1, 2, \ldots, 6$. for each $\boldsymbol{\Sigma}_i$, misclassification rates for the 9 classifiers were plotted against dimension, using three different colors (for different shrinkage methods) and three different line types (for different procedures), as described below.

Shrinkage towards **Identity** Matrix : **RED**
Shrinkage towards **Intra-class Correlation** Matrix : **GREEN**
Shrinkage towards **Diagonal** Matrix : **BLUE**

Minimization w.r.t shrinkage parameter : **DOTTED**
General Aggregation : **DASHED**
Case-dependent Aggregation : **DOT-DASHED**

Results and Discussion

First point to be noted is better performance of 'minimization w.r.t shrinkage' method compared to aggregation, because in most of the cases, the dotted line lies below the dashed or dot-dashed line. So, we do not gain much by using the more complex (computationally) aggregation procedures. Secondly, among the dotted lines, the red ones are more on the higher side compared to the blue or green ones. This suggests that shrinkage towards identity may not perform well, when the actual covariance matrix is substantially different from constant times identity. Thirdly, the classifier corresponding

to blue dotted lines perform more or less better in all the cases. This suggests shrinkage towards an appropriate diagonal matrix may be more useful in high dimensional classification problem. Lastly, in some cases the green dotted lines are comparable to the blue ones. Actually, there is a scope of improvement in the 'shrinkage towards Intra-class Correlation matrix' method, because in this method all the variances were assumed to be equal. Clearly, this assumption is not sufficient. So, if only the correlations of the variables are assumed to be equal with no restriction on the variances, the corresponding classifier may be a good competitor of the 'shrinkage towards diagonal' method.
So, the following question was raised after the simulation study.

- ***How much improvement in misclassification probability is possible if shrinkage towards a covariance matrix with possibly unequal variances and intra-class correlation structure is considered?***

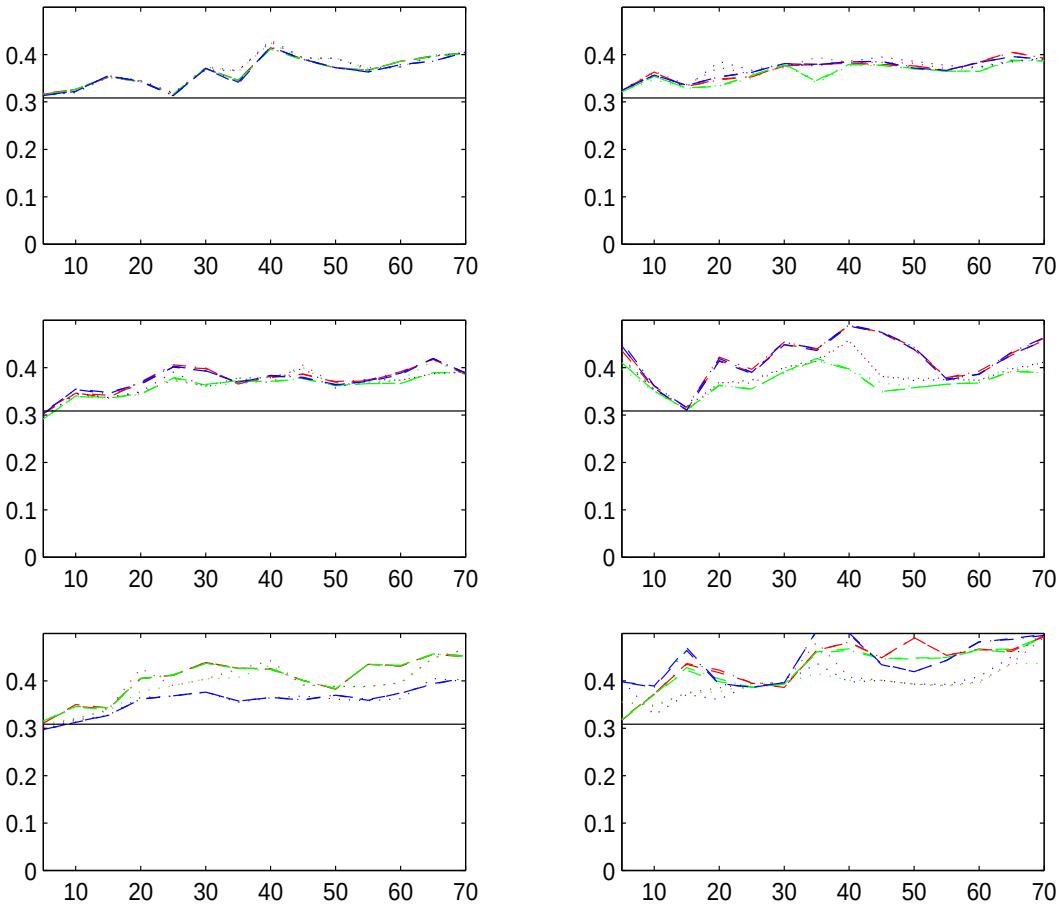


Figure 2.4.3: Misclassification Probability plotted against dimension for $N = 50$, $\Delta = 1$. Colors red, blue and green indicate shrinkage towards identity, intraclass correlation and diagonal matrices, respectively, while dotted, dashed and dot-dashed lines indicate pointwise minimum, general aggregation and case-specific aggregation, respectively. The black line shows the Bayes risk. The six panels (viewed row-wise) correspond to correct covariance matrices $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_6$, respectively.

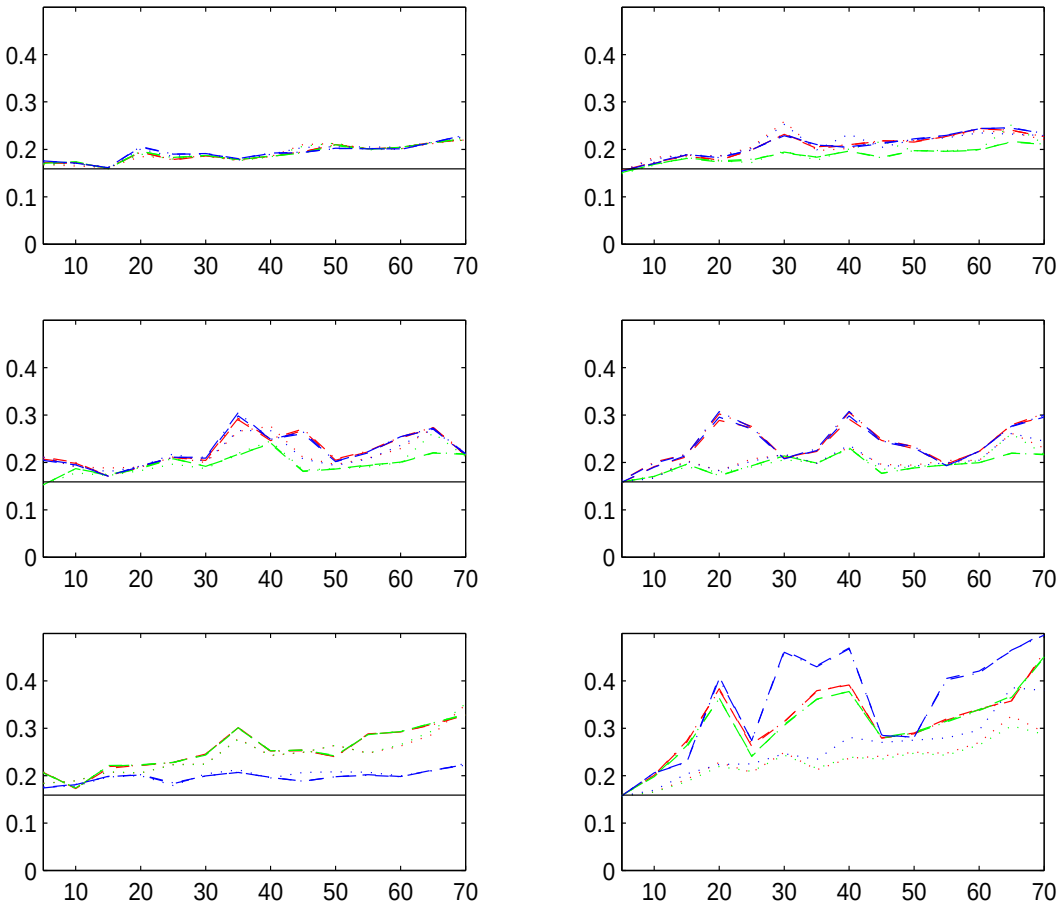


Figure 2.4.4: Misclassification Probability plotted against dimension for $N = 50$, $\Delta = 4$, $\mathbf{\Sigma} = \mathbf{I}$. Colors red, blue and green indicate shrinkage towards identity, intraclass correlation and diagonal matrices, respectively, while dotted, dashed and dot-dashed lines indicate pointwise minimum, general aggregation and case-specific aggregation, respectively. The black line shows the Bayes risk. The six panels (viewed row-wise) correspond to correct covariance matrices $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_6$, respectively.

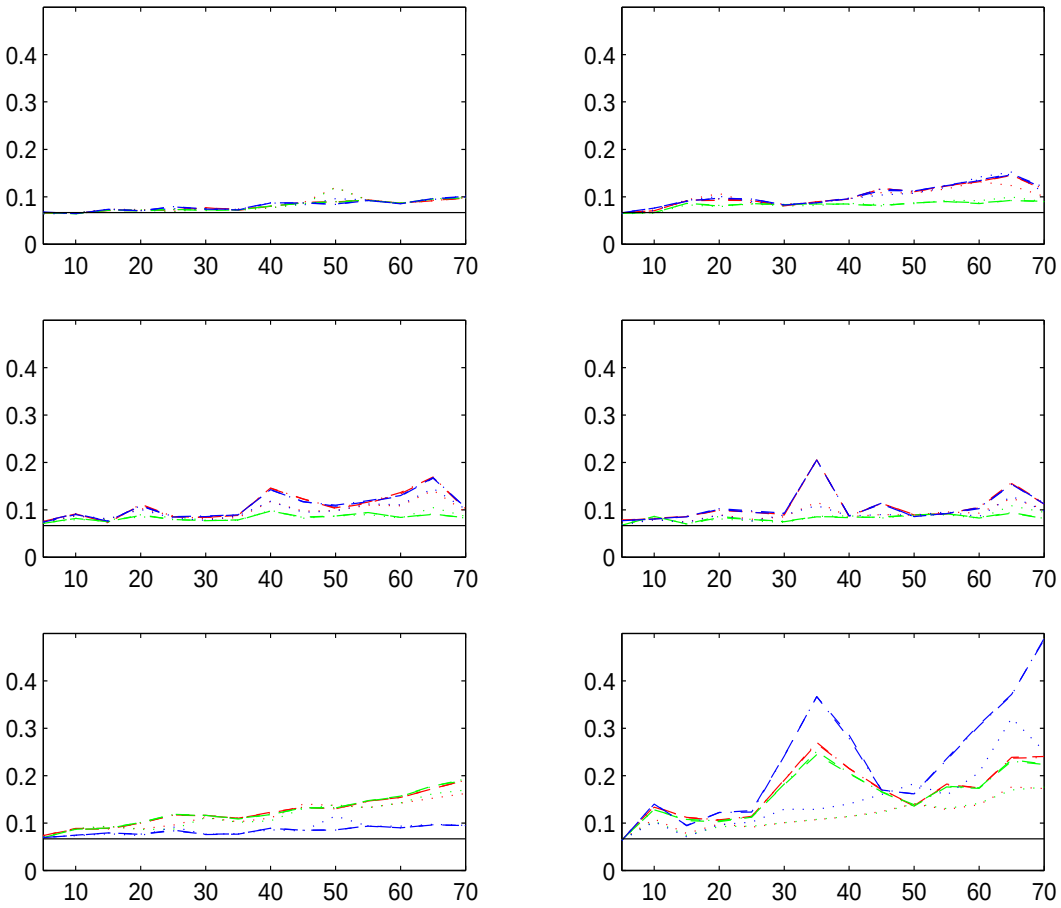


Figure 2.4.5: Misclassification Probability plotted against dimension for $N = 50$, $\Delta = 9$. Colors red, blue and green indicate shrinkage towards identity, intraclass correlation and diagonal matrices, respectively, while dotted, dashed and dot-dashed lines indicate pointwise minimum, general aggregation and case-specific aggregation, respectively. The black line shows the Bayes risk. The six panels (viewed row-wise) correspond to correct covariance matrices $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_6$, respectively.

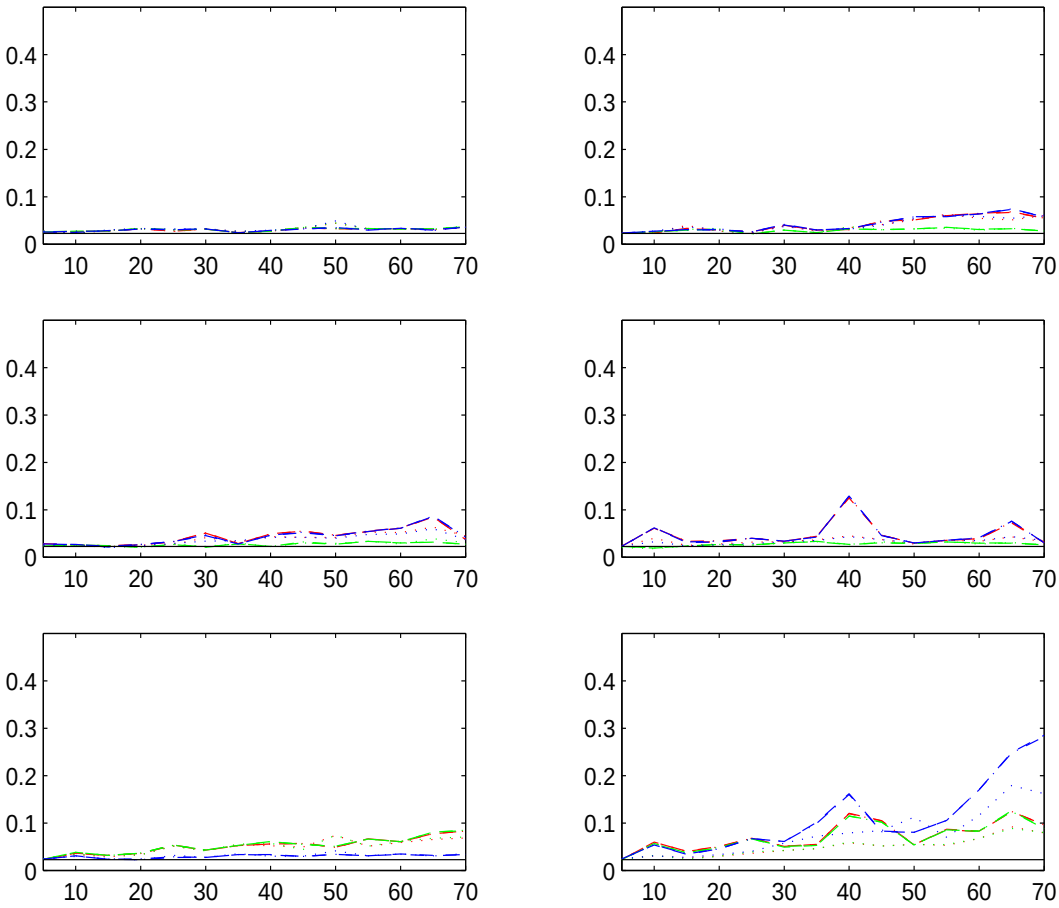


Figure 2.4.6: Misclassification Probability plotted against dimension for $N = 50$, $\Delta = 16$. Colors red, blue and green indicate shrinkage towards identity, intraclass correlation and diagonal matrices, respectively, while dotted, dashed and dot-dashed lines indicate pointwise minimum, general aggregation and case-specific aggregation, respectively. The black line shows the Bayes risk. The six panels (viewed row-wise) correspond to correct covariance matrices $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_6$, respectively.

2.5 Discussion of Simulation Results on Regularization

As it was realized from the simulation studies that shrinkage towards a suitable intraclass-correlation matrix with possibly different variances may generally improve the misclassification probability. Keeping this in mind, we tried to find the MLE of $\sigma_1^2, \ldots, \sigma_p^2$ and ρ under the assumption that$\mathbf{\Sigma} = \mathbf{DRD}$, where

$$\mathbf{D} = \begin{bmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_p \end{bmatrix}; \ \mathbf{R} = \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{bmatrix} = (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p.$$

This is equivalent to minimizing $\text{trace}(\mathbf{\Sigma}^{-1}\mathbf{S}) + \log(|\mathbf{\Sigma}|)$ w.r.t $\sigma_1, \ldots, \sigma_p, \rho$, where $\mathbf{S}$ is the sum of squares and product matrix

$$\mathbf{S} = \frac{\sum_{i=1}^{n}(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})^T + \sum_{i=1}^{n}(\mathbf{Y}_i - \overline{\mathbf{Y}})(\mathbf{Y}_i - \overline{\mathbf{Y}})^T}{2n}.$$

Now,

$$\begin{aligned}
&\text{trace}(\mathbf{\Sigma}^{-1}\mathbf{S}) + \ln|\mathbf{\Sigma}| \\
=&\text{trace}[\mathbf{D}^{-1}(a\mathbf{I}_p + b\mathbf{J}_p)\mathbf{D}^{-1}\mathbf{S}] + 2\ln|\mathbf{D}| + \ln|\mathbf{R}| \\
&\left(\text{where } a = \frac{1}{1-\rho}, b = -\frac{\rho}{(1-\rho)(1+\overline{p-1}\rho)} \right) \\
=&a.\text{trace}(\mathbf{D}^{-1}S\mathbf{D}^{-1}) + b.\text{trace}(\mathbf{1}_p^T\mathbf{D}^{-1}S\mathbf{D}^{-1}\mathbf{1}_p) + 2\ln|\mathbf{D}| + \ln|\mathbf{R}| \\
=&a.\sum_{i=1}^{p} S_{ii}\tau_i^2 + b.\tau^T\mathbf{S}\tau - 2\sum_{i=1}^{p}\ln\tau_i + \ln|\mathbf{R}| \ \left(\text{where } \tau = (\tau_1, \ldots, \tau_p) \text{ and } \tau_i = \frac{1}{\sigma_i} \right) \\
=&\tau^T A\tau - 2\sum_{i=1}^{p}\ln\tau_i + \ln|\mathbf{R}| \qquad (\text{where } A = a.\text{DIAG}(\mathbf{S}) + b.\mathbf{S}).
\end{aligned}$$

For fixed ρ, the above is minimized when$A\tau = 1/\tau$, where $1/\tau = (\tau_1^{-1}, \ldots, \tau_p^{-1})^T$. This is a nonlinear equation and has no closed form solution. So numerical methods were used (using $\mathbf{S}_{ii}$ as the initial estimate of σ_i) to minimize the function and to obtain the MLE.
Next, in order to compare this method with other shrinkage methods, a simulation study was conducted. There a surprising fact was observed. If the true $\mathbf{\Sigma}$ is known to have intra-class correlation structure even with different variances, the effect of error in mean-estimation on misclassification rates

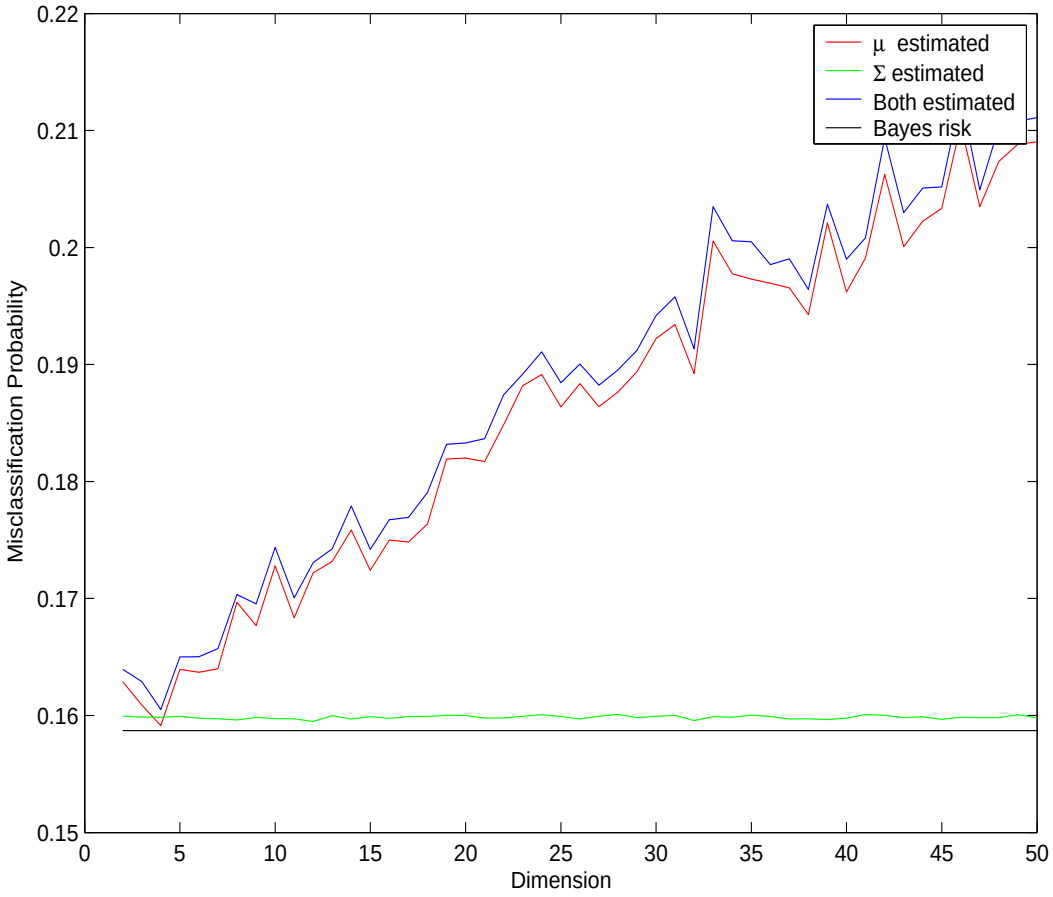


Figure 2.5.1: Plot of Bayes risk and misclassification probability of the Bayes classifiers under several restrictions on the parameters (based on a training sample of size 50), when the true populations are Gaussian with dispersion matrix having intraclass correlation structure

is much more than the effect of error in $\boldsymbol{\Sigma}$-estimation. The results of the simulation studies are displayed in Figure 2.5. This is surprising, because in high dimensional classification problems estimation of $\boldsymbol{\Sigma}$ is expected to be more critical than $\boldsymbol{\mu}$-estimation. The above simulation result gives rise to the following questions.

- ***What is more important in classification problem : $\boldsymbol{\mu}$-estimation, or $\boldsymbol{\Sigma}$-estimation?***
- ***To what extent the misclassification probability is affected, when $\boldsymbol{\mu}$ is known and $\boldsymbol{\Sigma}$ is estimated?***

These questions are addressed in the next chapter.

Chapter 3

Effects of Mean and Covariance Estimation

3.1 Comparison Between the Two Effects

In this section, we are trying to answer the first question, namely which of the two errors (corresponding to the estimation of the mean and the covariance matrix) contributes more to increase the misclassification probability. For this section, we will assume the sample size n to be large so that higher order terms (e.g., $O(n^{-2})$) can be ignored. Intuitively, the estimation of covariance matrix should be more important, as it involves more parameters than the mean. In this section, we have tried to give some theoretical justification to this intuitive reasoning.

Proposition 3.1.1. *In a classification problem with two populations $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, if sample means and sample covariance are used as estimates of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$, then the misclassification probability is approximately given by*

$$\textit{Bayes Risk} + \frac{1}{n}\frac{1}{2\sqrt{\Delta}}\left[\underbrace{(p-1)+\frac{\Delta}{4}}_{\substack{\textit{contribution}\\ \textit{due to } \boldsymbol{\mu}\textit{-estimation}}} + \underbrace{(p-1)\frac{\Delta}{4}}_{\substack{\textit{contribution}\\ \textit{due to } \boldsymbol{\Sigma}\textit{-estimation}}}\right]\phi\left(\frac{\sqrt{\Delta}}{2}\right)$$

.

Hence, contribution of error due to $\boldsymbol{\Sigma}$-estimation is more than that of means when $\Delta \geq 4(p-1)/(p-2) \approx 4$.

Proof. We know that, if the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are estimated by $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$

and $\hat{\boldsymbol{\Sigma}}$ respectively based on the training sample, then the conditional misclassification probability for a future observation $\mathbf{X} \sim \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ is given by

$$\begin{aligned}&\mathrm{P}[(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{X} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}) < 0] \\ &= \overline{\Phi}\left(\frac{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)}{\sqrt{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}}\right),\end{aligned}$$

where $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. So, the unconditional misclassification probability is given by

$$\mathrm{E}\left[\overline{\Phi}\left(\frac{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1 - (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2)}{\sqrt{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}}\right)\right]$$

while the Bayes risk is

$$\overline{\Phi}\left(\frac{\sqrt{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{2}\right) = \overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right).$$

By the symmetry of the problem, the unconditional misclassification probability for a future observation from $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ is the same as (3.1).
If $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$ and $\hat{\boldsymbol{\Sigma}}$ are taken to be the sample means and the pooled covariance, then the errors are of the order $n^{-1/2}$, where n is the common sample size of the training sample from the two population, as all these estimates are moment-estimates. So, let us assume that

$$\begin{aligned}\hat{\boldsymbol{\mu}}_1 &= \boldsymbol{\mu}_1 + \frac{\boldsymbol{\varepsilon}_1}{\sqrt{n}}, \\ \hat{\boldsymbol{\mu}}_2 &= \boldsymbol{\mu}_2 + \frac{\boldsymbol{\varepsilon}_2}{\sqrt{n}} \\ \text{and } \hat{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma} + \frac{\boldsymbol{\Lambda}}{\sqrt{n}}.\end{aligned}$$

Clearly, $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2 \sim \mathrm{N}(0, \boldsymbol{\Sigma})$ and $(n^{'}\hat{\boldsymbol{\Sigma}}) \sim W_p(n^{'}, \boldsymbol{\Sigma})$, where $n^{'} = 2n - 2$, and all of them are independent. Let

$$\begin{aligned}\boldsymbol{\delta}_1 &= \boldsymbol{\varepsilon}_1 - \boldsymbol{\varepsilon}_2 \\ \text{and } \boldsymbol{\delta}_2 &= \boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2\end{aligned}$$

Then, $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \sim N(0, 2\boldsymbol{\Sigma})$ and they are independent. Also, $\hat{\boldsymbol{\Sigma}}$ and hence $\boldsymbol{\Lambda}$ is independent of $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2$. Now, if we neglect higher order terms, then

$$
\begin{aligned}
&\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left(\boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2} \right) \\
&= \left(\boldsymbol{\mu}_1 + \frac{\varepsilon_1}{\sqrt{n}} - \boldsymbol{\mu}_2 - \frac{\varepsilon_2}{\sqrt{n}} \right)^T \left(\boldsymbol{\Sigma} + \frac{\boldsymbol{\Lambda}}{\sqrt{n}} \right)^{-1} \left(\boldsymbol{\mu}_1 - \frac{\boldsymbol{\mu}_1 + \frac{\varepsilon_1}{\sqrt{n}} + \boldsymbol{\mu}_2 + \frac{\varepsilon_2}{\sqrt{n}}}{2} \right) \\
&= \left(\boldsymbol{\mu} + \boldsymbol{\delta}_1/\sqrt{n} \right)^T \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(I + \frac{\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-\frac{1}{2}}}{\sqrt{n}} \right)^{-1} \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\frac{\boldsymbol{\mu} + \boldsymbol{\delta}_2/\sqrt{n}}{2} \right) \\
&\approx \frac{1}{2} \left[\left(\boldsymbol{\mu} + \frac{\boldsymbol{\delta}_1}{\sqrt{n}} \right)^T \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(I - \frac{\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-\frac{1}{2}}}{\sqrt{n}} + \frac{\boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-\frac{1}{2}}}{n} \right) \boldsymbol{\Sigma}^{-\frac{1}{2}} \left(\boldsymbol{\mu} + \frac{\boldsymbol{\delta}_2}{\sqrt{n}} \right) \right] \\
&= \frac{1}{2} \left[\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{1}{\sqrt{n}} \left(\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_1 + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_2 - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \right) \right] \\
&+ \frac{1}{2} \left[\frac{1}{n} \left(\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_1 - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_2 + \boldsymbol{\delta}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_2 \right) \right] \\
&= \frac{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}{2} \left[1 + \frac{a_1}{\sqrt{n}} + \frac{a_2}{n} \right] \qquad \text{(say)},
\end{aligned}
$$

where

$$
a_1 = \frac{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_1 + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_2 - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}
$$

and $a_2 = \dfrac{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_1 - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda} \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_2 + \boldsymbol{\delta}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\delta}_2}{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}}$.

Similarly,

$$
\begin{aligned}
&\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}} \\
&= \left(\boldsymbol{\mu}+\frac{\boldsymbol{\delta}_1}{\sqrt{n}}\right)^T\left(\boldsymbol{\Sigma}+\frac{\boldsymbol{\Lambda}}{\sqrt{n}}\right)^{-1}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+\frac{\boldsymbol{\Lambda}}{\sqrt{n}}\right)^{-1}\left(\boldsymbol{\mu}+\frac{\boldsymbol{\delta}_1}{\sqrt{n}}\right) \\
&= \left(\boldsymbol{\mu}+\frac{\boldsymbol{\delta}_1}{\sqrt{n}}\right)^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(I+\frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}}{\sqrt{n}}\right)^{-2}\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\mu}+\frac{\boldsymbol{\delta}_1}{\sqrt{n}}\right) \\
&\approx \left(\boldsymbol{\mu}+\frac{\boldsymbol{\delta}_1}{\sqrt{n}}\right)^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(I-2\frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}}{\sqrt{n}}+3\frac{(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}})^2}{n}\right)\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\mu}+\frac{\boldsymbol{\delta}_1}{\sqrt{n}}\right) \\
&= \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\frac{1}{\sqrt{n}}\left(2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)+ \\
&\qquad \frac{1}{n}\left(3\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-4\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1+\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_2\right) \\
&= \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\left[1+\frac{b_1}{\sqrt{n}}+\frac{b_2}{n}\right] \qquad \text{(say)},
\end{aligned}
$$

where

$$
\begin{aligned}
b_1 &= \frac{2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} \\
\text{and } b_2 &= \frac{3\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-4\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1+\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\frac{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1-(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)/2)}{\sqrt{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}} &\approx \frac{((\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})/2)\left[1+n^{-1/2}a_1+n^{-1}a_2\right]}{\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\left[1+n^{-1/2}b_1+n^{-1}b_2\right]}} \\
&= \frac{\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{2}\left[1+\frac{a_1}{\sqrt{n}}+\frac{a_2}{n}\right]\left[1+\frac{b_1}{\sqrt{n}}+\frac{b_2}{n}\right]^{-\frac{1}{2}} \\
&\approx \frac{\sqrt{\Delta}}{2}\left[1+\frac{a_1}{\sqrt{n}}+\frac{a_2}{n}\right]\left[1-\frac{b_1}{2\sqrt{n}}+\frac{1}{n}\left(-\frac{b_2}{2}+\frac{3}{8}b_1^2\right)\right] \\
&\approx \frac{\sqrt{\Delta}}{2}\left[1+\frac{1}{\sqrt{n}}\left(a_1-\frac{b_1}{2}\right)+\frac{1}{n}\left(a_2-\frac{a_1b_1}{2}-\frac{b_2}{2}+\frac{3}{8}b_1^2\right)\right] \\
&= \frac{\sqrt{\Delta}}{2}\left[1+\frac{c_1}{\sqrt{n}}+\frac{c_2}{n}\right] \qquad \text{(say)}
\end{aligned}
$$

Using the Taylor-Expansion for the function $\overline{\Phi}(.)$ about $\frac{\sqrt{\Delta}}{2}$ up to the first

order, we get

$$\begin{aligned}
&\overline{\Phi}\left(\frac{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1-(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)/2)}{\sqrt{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}}\right)\\
&\quad\approx\overline{\Phi}\left[\frac{\sqrt{\Delta}}{2}\left(1+\frac{c_1}{\sqrt{n}}+\frac{c_2}{n}\right)\right]\\
&\quad\approx\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{\sqrt{\Delta}}{2}\left(\frac{c_1}{\sqrt{n}}+\frac{c_2}{n}\right)\left[-\phi\left(\frac{\sqrt{\Delta}}{2}\right)\right]+\frac{1}{2}\frac{\Delta}{4}\frac{c_1^2}{n}\left[-\phi'\left(\frac{\sqrt{\Delta}}{2}\right)\right]\\
&\qquad\text{(as we are interested in the terms of order } \frac{1}{n}\text{)}.
\end{aligned}$$

So, the misclassification probability can be approximated as

$$\begin{aligned}
&\mathrm{E}\left[\overline{\Phi}\left(\frac{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1-(\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2)/2)}{\sqrt{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}}\right)\right]\\
&\quad\approx\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{\sqrt{\Delta}}{2}\left(\frac{\mathrm{E}(c_1)}{\sqrt{n}}+\frac{\mathrm{E}(c_2)}{n}\right)\left[-\phi\left(\frac{\sqrt{\Delta}}{2}\right)\right]\\
&\qquad+\frac{1}{n}\frac{\Delta}{8}\,\mathrm{E}c_1^2\left[-\phi'\left(\frac{\sqrt{\Delta}}{2}\right)\right].
\end{aligned}$$

Now,

$$\mathrm{E}(c_1)=\mathrm{E}\left(a_1-\frac{b_1}{2}\right)=\mathrm{E}\left[\frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_2}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right]=0\qquad\qquad(\text{as } \mathrm{E}(\boldsymbol{\delta}_2)=0),$$

and

$$\begin{aligned}
\mathrm{E}(c_2)=&\mathrm{E}\left[a_2-\frac{b_2}{2}-\frac{a_1b_1}{2}+\frac{3}{2}\left(\frac{b_1}{2}\right)^2\right]\\
=&\mathrm{E}\left[\frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta}_1+\boldsymbol{\delta}_2)+\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_2}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right]\\
&-\mathrm{E}\left[\frac{\frac{3}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1+\frac{1}{2}\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right]\\
&-\mathrm{E}\left[\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_2)(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})}{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}\right]\\
&+\frac{3}{2}\left[\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}\right]
\end{aligned}$$

$$=\mathrm{E}\left[\frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{\Delta}\right]-\mathrm{E}\left[\frac{\frac{3}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\frac{1}{2}\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1}{\Delta}\right]$$

$$-\mathrm{E}\left[\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}{\Delta^2}\right]+\frac{3}{2}\mathrm{E}\left[\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}{\Delta^2}\right]$$

(as $\mathrm{E}(\boldsymbol{\delta}_1)=\mathrm{E}(\boldsymbol{\delta}_2)=\mathbf{0}$, $\mathrm{E}(\boldsymbol{\Lambda})=\mathbf{0}$, and $\boldsymbol{\delta}_1,\boldsymbol{\delta}_2$ and $\boldsymbol{\Lambda}$ are independent)

$$=\mathrm{E}\left[\frac{-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1}{\Delta}\right]+\frac{1}{2}\mathrm{E}\left[\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}{\Delta^2}\right]$$

$$=\mathrm{E}\left[\frac{-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1}{\Delta}\right]+\frac{1}{2}\mathrm{E}\left[\frac{(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1)^2+(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}{\Delta^2}\right]$$

(as $\boldsymbol{\delta}_1$ and $\boldsymbol{\Lambda}$ are independent)

$$=-\frac{1}{2}\frac{\Delta\mathrm{E}(\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1)-\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1)^2+\Delta\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})-\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2}{\Delta^2}$$

$$=-\frac{2p\Delta-2\Delta+\frac{n}{n'}(p+1)\Delta^2-\frac{n}{n'}2\Delta^2}{2\Delta^2}$$

(as $\boldsymbol{\delta}_1\sim N(\mathbf{0},2\boldsymbol{\Sigma})$, $\dfrac{\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1}{2}\sim\chi^2_p$,

we have, $\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1)^2=\mathrm{Var}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1)=2\Delta$,$\mathrm{E}(\boldsymbol{\delta}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_1)=2p$)

(The other two quantities are obtained from Eq(1))

$$=-\frac{p-1}{2\Delta^2}[2\Delta+\frac{n}{n'}\Delta^2],$$

because

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2\\
=&n\mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}}{\sqrt{n}}\right)\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]^2\\
=&n\mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-\frac{1}{2}}-I\right)\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]^2\\
=&n\mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\frac{\mathbf{S}}{n^{'}}\boldsymbol{\Sigma}^{-\frac{1}{2}}-I\right)\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]^2 \qquad (\text{as } \hat{\boldsymbol{\Sigma}}=\frac{\mathbf{S}}{n^{'}}=\boldsymbol{\Sigma}+\frac{\boldsymbol{\Lambda}}{\sqrt{n}})\\
=&n(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2\mathrm{E}\left[\frac{1}{n^{'}}\frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{W}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}-1\right]^2\\
&\qquad (\mathbf{W}=\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{S}\boldsymbol{\Sigma}^{-\frac{1}{2}}\sim W_p(n^{'},I), \text{ as } \mathbf{S}\sim W_p(n^{'},\boldsymbol{\Sigma}))\\
=&n\Delta^2\mathrm{E}\left[\frac{Y}{n^{'}}-1\right]^2 \qquad (\ Y=\frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{W}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\sim\chi^2_{n^{'}})\\
=&n\Delta^2\left[\frac{\mathrm{E}Y^2}{n^{'2}}+1-2\frac{\mathrm{E}Y}{n^{'}}\right]\\
=&n\Delta^2\left[\frac{n^{'2}+2n^{'}}{n^{'2}}+1-2\frac{n^{'}}{n^{'}}\right]=2\frac{n}{n^{'}}\Delta^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\\
=&n\mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\frac{\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-\frac{1}{2}}}{\sqrt{n}}\right)^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]\\
=&n\mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\frac{\mathbf{W}}{n^{'}}-I\right)^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right] \qquad (\text{where } \mathbf{W} \text{ is as above, } \mathbf{W}\sim W_p(n^{'},I))\\
=&n\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\left(\frac{\mathbf{W}}{n^{'}}-I\right)^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]=n\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\frac{\mathrm{E}\mathbf{W}^2}{n^{'2}}+I-2\frac{\mathrm{E}\mathbf{W}}{n^{'}}\right)^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]\\
=&n\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\frac{n^{'}(n^{'}+p+1)}{n^{'2}}I+I-2\frac{n^{'}}{n^{'}}I\right)^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]=(p+1)\frac{n}{n^{'}}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})=(p+1)\frac{n}{n^{'}}\Delta.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\mathrm{E}c_1^2=&\mathrm{E}\left[a_1-\frac{b_1}{2}\right]^2=\mathrm{E}\left[\frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_2}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}\right]^2 \qquad (\text{substituting the values of } a_1 \text{ and } b_1)\\
=&\frac{\mathrm{Var}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}_2)}{\Delta^2}=\frac{2\Delta}{\Delta^2} \qquad (\text{as } \boldsymbol{\delta}_2\sim N(0,2\boldsymbol{\Sigma}))
\end{aligned}
$$

Hence, collecting all the coefficients, we have the misclassification probability

$$
\begin{aligned}
&\mathrm{E}\left[\overline{\Phi}\left(\frac{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_1-\frac{\hat{\boldsymbol{\mu}}_1+\hat{\boldsymbol{\mu}}_2}{2})}{\sqrt{\hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}}}\right)\right] \\
\approx&\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{1}{n}\left[\frac{\sqrt{\Delta}}{2}\frac{p-1}{2\Delta^2}(2\Delta+\frac{n}{n'}\Delta^2)\phi\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{\Delta}{8}\frac{2\Delta}{\Delta^2}\left\{-\phi'\left(\frac{\sqrt{\Delta}}{2}\right)\right\}\right] \\
=&\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{1}{n}\left[\frac{\sqrt{\Delta}}{2}\frac{p-1}{2\Delta^2}(2\Delta+\frac{n}{n'}\Delta^2)\phi\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{\Delta}{8}\frac{2\Delta}{\Delta^2}\frac{\sqrt{\Delta}}{2}\phi\left(\frac{\sqrt{\Delta}}{2}\right)\right] \\
&(\text{as } \phi'(x)=-x\phi(x)) \\
=&\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{1}{n}\frac{\sqrt{\Delta}}{2\Delta^2}\left[(p-1)\Delta+\frac{\Delta^2}{4}+(p-1)\frac{n}{2n-2}\frac{\Delta^2}{2}\right]\phi\left(\frac{\sqrt{\Delta}}{2}\right) \\
=&\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{1}{n}\frac{1}{2\sqrt{\Delta}}\left[(p-1)+\frac{\Delta}{4}+(p-1)\frac{n}{2n-2}\frac{\Delta}{2}\right]\phi\left(\frac{\sqrt{\Delta}}{2}\right) \\
\approx&\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)+\frac{1}{n}\frac{1}{2\sqrt{\Delta}}\left[\underbrace{(p-1)+\frac{\Delta}{4}}_{\substack{\text{contribution}\\ \text{due to }\boldsymbol{\mu}\text{-estimation}}}+\underbrace{(p-1)\frac{\Delta}{4}}_{\substack{\text{contribution}\\ \text{due to }\boldsymbol{\Sigma}\text{-estimation}}}\right]\phi\left(\frac{\sqrt{\Delta}}{2}\right)
\end{aligned}
$$

Clearly, the contribution of the error of $\boldsymbol{\Sigma}$-estimation is more, when $(p-1)\Delta/4 \geq (p-1)+\Delta/4$ or equivalently $\Delta \geq 4(p-1)/(p-2) \approx 4$. □

However, when the covariance matrix is known to have intra-class correlation structure with possibly different variances, the estimation of the means turns out to be more important. The simulation study shows that, in such a situation, if $\boldsymbol{\mu}$ is known, the misclassification probability is quite close to the Bayes risk [as shown in Figure 2.5]. On the other hand, if $\boldsymbol{\mu}$ is unknown, even when $\boldsymbol{\Sigma}$ is known, the misclassification probability may be quite far away from the Bayes risk. To prove this phenomenon theoretically is difficult, though a special case of this has been discussed in the next proposition. One possible reason for this is that the number of parameters in $\boldsymbol{\Sigma}$ becomes small when $\boldsymbol{\Sigma}$ has intra-class correlation structure.

Proposition 3.1.2. *If all the correlations are ignored and the variances are estimated by the corresponding sample analogues, then the misclassification*

probability is approximately given by

$$\textit{Misclassification Probability} = \textit{Bayes Risk} + \underbrace{[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T\mathbf{R}^{-1}\mathbf{l} - (\mathbf{l}^T l)^2]\frac{\phi\left[\frac{\sqrt{\Delta}}{2}\right]}{2\Delta\sqrt{\Delta}}}_{\textit{Bias for using wrong structure of } \mathbf{\Sigma}} + \underbrace{\frac{1}{n-1}\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}r_{ij}^2))\mathbf{l} - \sum_{i,j}\mathbf{l}_i^2\mathbf{l}_j^2 r_{ij}^2\right]\frac{\phi\left[\frac{\sqrt{\Delta}}{2}\right]}{2\Delta\sqrt{\Delta}}}_{n^{-1} \textit{ order term}},$$

where $\mathbf{l} = \mathbf{R}^{-1}\mathbf{D}^{-1}\boldsymbol{\mu}$. *Here* $\mathbf{R}$ *is the true correlation matrix,* $\mathbf{R}^{-1} = ((r^{ij}))$ *and* $\mathbf{\Sigma} = \mathbf{DRD}$, *where* $\mathbf{D}$ *is a diagonal matrix with the standard deviations as the diagonal entries.*

Proof. As we have seen in the earlier proposition, if error in estimation of $\mathbf{\Sigma}$ is of the order $n^{-1/2}$ i.e., $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma} + \mathbf{\Lambda}n^{-1/2}$, then the resulting misclassification probability is approximately equal to

$$\underbrace{\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)}_{\textit{Bayes Risk}} + \frac{1}{n}\frac{\sqrt{\Delta}}{2}\left[\frac{\Delta\mathrm{E}(\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}) - \mathrm{E}(\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\boldsymbol{\mu})^2}{\Delta^2}\right]\phi\left[\frac{\sqrt{\Delta}}{2}\right] \tag{3.1.1}$$

Now, if all the correlations are ignored and the variances are estimated from the training data, then the natural estimate to consider is

$$\hat{\mathbf{\Sigma}}_D = \begin{bmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\sigma}_p^2 \end{bmatrix},$$

where

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^n (X_{ik} - \overline{X}_k)^2 + \sum_{i=1}^n (Y_{ik} - \overline{Y}_k)^2}{2(n-1)}; \quad , k = 1, 2, \dots, p.$$

Here, $\mathbf{X}_i = (X_{i1}, \dots, X_{ip}); i = 1, \dots, n$ and $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip}); i = 1, \dots, n$ are the training samples.

Now we need to calculate

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})\\
&\quad = n\ \mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\frac{\boldsymbol{\Lambda}}{\sqrt{n}}\boldsymbol{\Sigma}^{-1}\frac{\boldsymbol{\Lambda}}{\sqrt{n}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]\\
&\quad = n\ \mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\frac{\boldsymbol{\Lambda}}{\sqrt{n}}\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]\\
&\quad = n\ \mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left\{\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\hat{\boldsymbol{\Sigma}}_D-\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-\frac{1}{2}}\right\}^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]\\
&\quad = n\ \mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-\frac{1}{2}}-\mathbf{I}\right)^2\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\right]\\
&\quad = n\ \mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2\\
&\quad = n\ \mathrm{E}\left(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\frac{\boldsymbol{\Lambda}}{\sqrt{n}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)^2\\
&\quad = n\ \mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\Sigma}}_D-\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]^2\\
&\quad = n\ \mathrm{E}\left[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]^2\\
&\quad = n\ \mathrm{E}\left[(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2+(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}.\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right]
\end{aligned}
$$

So,

$$
\begin{aligned}
&\Delta\ \mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})-\mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2\\
&= n[\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\ \mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})-\ \mathrm{E}(\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}_D\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})^2]\\
&= n[\mathbf{l}^T\mathbf{R}\mathbf{l}\ \mathrm{E}\{\mathbf{l}^T(\mathbf{D}^{-1}\hat{\boldsymbol{\Sigma}}_D\mathbf{D}^{-1})\mathbf{R}^{-1}(\mathbf{D}^{-1}\hat{\boldsymbol{\Sigma}}_D\mathbf{D}^{-1})\mathbf{l}\}-\ \mathrm{E}\{\mathbf{l}^T(\mathbf{D}^{-1}\hat{\boldsymbol{\Sigma}}_D\mathbf{D}^{-1})\mathbf{l}\}^2]\\
&\qquad\qquad (\text{putting } \boldsymbol{\Sigma}^{-1}=\mathbf{D}^{-1}\mathbf{R}^{-1}\mathbf{D}^{-1} \text{ and } \mathbf{l}=\mathbf{R}^{-1}\mathbf{D}^{-1}\boldsymbol{\mu})\\
&= \varepsilon \qquad (\text{say})
\end{aligned}
$$

Now, $\mathbf{D}^{-1}\hat{\boldsymbol{\Sigma}}_D\mathbf{D}^{-1}=\mathbf{W}$(say) is nothing but

$$
\mathbf{W}=\begin{bmatrix}\frac{\hat{\sigma}_1^2}{\sigma_1^2} & 0 & \dots & 0\\ 0 & \frac{\hat{\sigma}_2^2}{\sigma_2^2} & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \frac{\hat{\sigma}_p^2}{\sigma_p^2}\end{bmatrix}
$$

Clearly, $2(n-1)W_{ii} \sim \chi^2(2n-2)\forall i$. So, $\mathrm{E}(W_{ii}) = 1$ and $\mathrm{E}(W_{ii}W_{jj})$ is given by

$$\mathrm{E}(W_{ii}W_{jj}) = \mathrm{E}\left(\frac{\hat{\sigma}_i^2}{\sigma_i^2}\frac{\hat{\sigma}_j^2}{\sigma_j^2}\right) = 1 + \frac{r_{ij}^2}{n-1}\forall i,j = 1,2,\ldots,p$$

Hence, we have

$$\begin{aligned}
\varepsilon =& n[\mathbf{l}^T\mathbf{R}\mathbf{l}\ \mathrm{E}(\mathbf{l}^T\mathbf{W}\mathbf{R}^{-1}\mathbf{W}\mathbf{l}) - \ \mathrm{E}(\mathbf{l}^T\mathbf{W}\mathbf{l})^2] \\
=& n\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T\ \mathrm{E}((r^{ij}W_{ii}W_{jj}))\mathbf{l} - \ \mathrm{E}\left(\sum_{i=1}^{p} l_i^2 W_{ii}\right)^2\right] \\
=& n\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}\ \mathrm{E}W_{ii}W_{jj}))\mathbf{l} - \left(\sum_{i,j=1}^{p} l_i^2 l_j^2\ \mathrm{E}W_{ii}W_{jj}\right)^2\right] \\
=& n\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T\left(\left(r^{ij}\overline{1 + \frac{r_{ij}^2}{n-1}}\right)\right)\mathbf{l} - \left(\sum_{i,j=1}^{p} l_i^2 l_j^2\overline{1 + \frac{r_{ij}^2}{n-1}}\right)^2\right] \\
=& n\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}))\mathbf{l} - \left(\sum_{i,j=1}^{p} l_i^2 l_j^2\right) + \frac{1}{n-1}\left\{\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}r_{ij}^2))\mathbf{l} - \sum_{i,j=1}^{p} l_i^2 l_j^2 r_{ij}^2\right\}\right] \\
=& n\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T\mathbf{R}^{-1}\mathbf{l} - (\mathbf{l}^T\mathbf{l})^2 + \frac{1}{n-1}\left\{\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}r_{ij}^2))\mathbf{l} - \sum_{i,j=1}^{p} l_i^2 l_j^2 r_{ij}^2\right\}\right].
\end{aligned}$$

Plugging in the value of ε in (3.1.1), we get

misclassification probability

$$\begin{aligned}
&\approx \overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right) + \frac{1}{n}\frac{\sqrt{\Delta}}{2\Delta^2}\left[n\{\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T\mathbf{R}^{-1}\mathbf{l} - (\mathbf{l}^T\mathbf{l})^2\} + \left\{\frac{n}{n-1}\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}r_{ij}^2))\right.\right. \\
&\qquad\qquad \left.\left. - \frac{n}{n-1}\sum_{i,j=1}^{p} l_i^2 l_j^2 r_{ij}^2\right\}\right]\phi\left[\frac{\sqrt{\Delta}}{2}\right] \\
&= \underbrace{\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right)}_{\textit{Bayes Risk}} + \underbrace{[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T\mathbf{R}^{-1}\mathbf{l} - (\mathbf{l}^T\mathbf{l})^2]\frac{\phi\left[\frac{\sqrt{\Delta}}{2}\right]}{2\Delta\sqrt{\Delta}}}_{\textit{Bias for assuming wrong } \boldsymbol{\Sigma}}
\end{aligned}$$

$$\underbrace{+\frac{1}{n-1}\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}r_{ij}^2))\mathbf{l}-\sum_{i,j=1}^{p}l_i^2l_j^2r_{ij}^2\right]\frac{\phi\left[\frac{\sqrt{\Delta}}{2}\right]}{2\Delta\sqrt{\Delta}}}_{1/n\text{-order term}}$$

This completes the decomposition of misclassification probability into different parts. □

Following corollary shows the advantage of ignoring all correlations when the true $\mathbf{\Sigma}$ is a diagonal matrix,

Corollary 3.1.3. *When the true $\mathbf{\Sigma}$ is diagonal, i.e the correlation matrix $\mathbf{R}=\mathbf{I}_p$, then there is no bias term for ignoring all correlations, and it is better to use diagonal matrix rather than sample covariance matrix.*

Proof. The bias term for ignoring the correlations is $\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T\mathbf{R}^{-1}\mathbf{l}-(\mathbf{l}^T\mathbf{l})^2=0$, where $\mathbf{l}$ is as in Proposition 3.1.2. Also,

$1/n$-order error term in misclassification probability, when $\hat{\mathbf{\Sigma}}_D$ is used

$$\frac{c(\Delta)}{n-1}\left[\mathbf{l}^T\mathbf{R}\mathbf{l}.\mathbf{l}^T((r^{ij}r_{ij}^2))\mathbf{l}-\sum_{i,j=1}^{p}l_i^2l_j^2r_{ij}^2\right]$$

$$=\frac{c(\Delta)}{n-1}\left[(\mathbf{l}^T\mathbf{l})^2-\sum_{i=1}^{p}l_i^4\right]\qquad(\text{as }\mathbf{R}=\mathbf{I}_p, r_{ij}=r_{ij}^2=r^{ij}=\boldsymbol{\delta}_{ij})$$

$$\leq\frac{c(\Delta)}{n-1}\left[(\mathbf{l}^T\mathbf{l})^2-\frac{1}{p}(\mathbf{l}^T\mathbf{l})^2\right]\qquad\left[\text{by Cauchy-Schwartz Inequality }\left(\sum_{i=1}^{p}l_i^2\right)^2\leq p\sum_{i=1}^{p}l_i^4\right]$$

$$=\frac{c(\Delta)}{n-1}\frac{p-1}{p}(\mathbf{l}^T\mathbf{l})^2$$

$$\leq\frac{c(\Delta)}{2(n-1)}(p-1)(\mathbf{l}^T\mathbf{l})^2$$

$$=\frac{c(\Delta)}{2(n-1)}(p-1)\Delta^2\qquad(\text{as }\Delta=\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}=\boldsymbol{\mu}^TD^{-2}\boldsymbol{\mu}=\mathbf{l}^T\mathbf{l})$$

$=1/n$-order error term in misclassification probability, when $\hat{\mathbf{\Sigma}}$ is used.

Hence, misclassification probability for using the diagonal matrix is smaller than that of using the sample covariance matrix. □

So, in general the contribution of error due to covariance estimation to the misclassification probability is more than that for mean estimation, unless

the populations are quite close to each other or the covariance $\boldsymbol{\Sigma}$ is highly parameterized. So, in the following sections, we will assume that mean-separation vector $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ to be known to study the effect of $\hat{\boldsymbol{\Sigma}}$ in the misclassification probability and whether it can be improved using regularization.

3.2 Effect of Estimation of $\boldsymbol{\Sigma}$ in Classification Problems

In this section, we will study the extent, to which the future misclassification probability can be affected, when an estimate of $\boldsymbol{\Sigma}$ is used. In this study, since we are concentrating on the effect of $\boldsymbol{\Sigma}$-estimation, we are assuming that the mean-separation vector is known for time being.

3.2.1 Nonsingular $\hat{\boldsymbol{\Sigma}}$

So, let us consider two populations $\mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with known mean-separation vector $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ and unknown $\boldsymbol{\Sigma}$, which is estimated by $\hat{\boldsymbol{\Sigma}}$. This $\hat{\boldsymbol{\Sigma}}$ can also be considered as a "wrong" value of $\boldsymbol{\Sigma}$.
The Bayes Risk of this problem is given by

$$\overline{\Phi}\left(\frac{\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{2}\right) = \overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right),$$

where $\Delta = \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ represents the Mahalanobis distance between the two populations. Misclassification probability for using $\hat{\boldsymbol{\Sigma}}$ in place of $\boldsymbol{\Sigma}$ is

$$\overline{\Phi}\left(\frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}}{2\sqrt{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}}}\right) = \overline{\Phi}\left(\frac{\sqrt{\Delta_*}}{2}\right).$$

Then, we have the following proposition.

Proposition 3.2.1.

$$\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right) \leq \overline{\Phi}\left(\frac{\sqrt{\Delta_*}}{2}\right) \leq \overline{\Phi}\left(\frac{2\sqrt{k}}{1+k}\frac{\sqrt{\Delta}}{2}\right),$$

where $k = \frac{\lambda_{max}(\mathbf{R})}{\lambda_{min}(\mathbf{R})}$ *and* $\mathbf{R}$ *is the ratio matrix* $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$. *Moreover, the lower bound is attained when* $\boldsymbol{\mu}$ *is any eigenvector of* $\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}$, *whereas the upper*

bound is attained when $\boldsymbol{\mu} = c(\mathbf{u} + \mathbf{v})$*, where* $\mathbf{u}$ *and* $\mathbf{v}$ *are two eigenvectors of* $\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}$ *corresponding to the largest and the smallest eigenvalues.*

Proof. To compare the Bayes risk and the future misclassification probability, we consider the ratio of the corresponding arguments of the $\overline{\Phi}$ function.

$$\begin{aligned}\frac{\frac{\sqrt{\Delta}}{2}}{\frac{\sqrt{\Delta_*}}{2}} =& \frac{\sqrt{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}}\sqrt{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}}{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}} \\ =& \frac{\sqrt{\nu^T\mathbf{R}\nu}\sqrt{\nu^T\mathbf{R}^{-1}\nu}}{\nu^T\nu} \\ =& r \qquad \text{(say)},\end{aligned}$$

where $\nu = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\mu}$ and $\mathbf{R} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$. Now, by Kantorovich inequality

$$r \leq \frac{\lambda_{max}(\mathbf{R}) + \lambda_{min}(\mathbf{R})}{2\sqrt{\lambda_{max}(\mathbf{R})\lambda_{min}(\mathbf{R})}} = \frac{1+k}{2\sqrt{k}},$$

where $k = \lambda_{max}(\mathbf{R})/\lambda_{min}(\mathbf{R})$. Again, by Cauchy-Schwartz inequality,
$r \geq 1$. To see this, consider the spectral decomposition of $\mathbf{R}$. Let

$$\begin{aligned}\mathbf{R} &= P\boldsymbol{\Lambda}P^T \\ \nu &= P\boldsymbol{\alpha},\end{aligned}$$

where $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \ldots \geq \lambda_p$ and $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_p)$. Then

$$r^2 = \frac{\sum_{i=1}^p \lambda_i\alpha_i^2 \sum_{i=1}^p \frac{1}{\lambda_i}\alpha_i^2}{\left(\sum_{i=1}^p \alpha_i^2\right)^2} \geq 1 \quad \text{(by Cauchy Schwartz Inequality)}.$$

Combining (3.2.1) and (3.2.1)

$$1 \leq r \leq \frac{1+k}{2\sqrt{k}}.$$

Hence

$$\overline{\Phi}\left(\frac{\sqrt{\Delta}}{2}\right) \leq \overline{\Phi}\left(\frac{\sqrt{\Delta_*}}{2}\right) \leq \overline{\Phi}\left(\frac{2\sqrt{k}}{1+k}\frac{\sqrt{\Delta}}{2}\right).$$

It is important to note that both bounds of r are attainable for suitable $\boldsymbol{\mu}$'s. For example, if $\nu = c(\mathbf{u} + \mathbf{v})$, where c is any constant, $\mathbf{u}$ and $\mathbf{v}$ are two eigenvectors of $\mathbf{R}$ corresponding to λ_1 and λ_p respectively (or equivalently, $\boldsymbol{\mu} = c(\mathbf{u} + \mathbf{v})$, where c is any constant, $\mathbf{u}$ and $\mathbf{v}$ are two eigenvectors of $\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}$ corresponding to λ_1 and λ_p respectively), then $r = (1+k)/2\sqrt{k}$. On the other hand, when ν is an eigenvector of $\mathbf{R}$ (or equivalently, $\boldsymbol{\mu}$ is an eigenvector of $\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}$), then $r = 1$. □

So, we never gain by using a "wrong" $\boldsymbol{\Sigma}$, unlike what happens if "wrong" $\boldsymbol{\mu}$ is used. The amount of loss heavily depends on the orientation of the mean-separation vector $\boldsymbol{\mu}$ w.r.t the ratio matrix $\mathbf{R}$. The worst possible misclassification probability depends on k, which acts as a measure of error for using $\hat{\boldsymbol{\Sigma}}$ instead of $\boldsymbol{\Sigma}$, and it approaches $\frac{1}{2}$ as $k \to \infty$ irrespective of Δ. If $\mathbf{R}$ has many distinct pairs of eigenvalues with high ratio, for each of the pairs there will be a possible direction of $\boldsymbol{\nu}$ leading to high misclassification probability. The ideal case is $k = 1$, when $\hat{\boldsymbol{\Sigma}}$ needs to be a multiple of $\boldsymbol{\Sigma}$.
On the other hand, if $\boldsymbol{\nu}$ happens to be an eigenvector of $\mathbf{R}$, or equivalently $\boldsymbol{\mu}$ turns out to be an eigenvector of $\hat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}$ (or equivalently of $\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}$), we do not loose at all. So, if $\mathbf{R}$ has large eigenvalues with high multiplicity, then it is more likely to have low misclassification probability. In the ideal case, $\mathbf{R}$ has only one distinct eigenvalue, which is possible if $\hat{\boldsymbol{\Sigma}}$ is a multiple of $\boldsymbol{\Sigma}$.

3.2.2 Some special Cases

Now, let us have a look at some special cases of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$.

1. When $\boldsymbol{\Sigma}$ has the intraclass correlation structure with correlation ρ, and $\hat{\boldsymbol{\Sigma}} = \text{Diag}(\boldsymbol{\Sigma})$, then $k = (1 + p\frac{\rho}{1-\rho})$ is the condition number of the true correlation matrix, which goes to ∞ as the correlation approaches 1 or $p \to \infty$. We know that any contrast and the vector of ones are eigenvectors of the correlation matrix. Now, the mean-separation vector $\boldsymbol{\mu}$ can be written as

$$\boldsymbol{\mu} = \sqrt{p}\overline{\boldsymbol{\mu}}\left(\frac{1}{\sqrt{p}}\mathbf{1}\right) + \sqrt{\sum_{i=1}^{p}(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^2}\,(\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}\mathbf{1})\,.$$

   So, the worst $\boldsymbol{\mu}$ is that for which $\sqrt{p}\overline{\boldsymbol{\mu}} = \sqrt{\sum_{i=1}^{p}(\boldsymbol{\mu}_i - \overline{\boldsymbol{\mu}})^2}$ or equivalently coefficient of variation of the vector $\boldsymbol{\mu}$ is 1. Performance of $\text{Diag}(\boldsymbol{\Sigma})$ improves as the coefficient of variation goes away from 1 in either direction.

2. When both $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ have intra-class correlation structure with correlations ρ and $\hat{\rho}$ respectively, then

$$\begin{aligned}\boldsymbol{\Sigma} &= (1-\rho)\mathbf{I}_p + \rho\mathbf{1}\mathbf{1}^T \\ \text{and } \hat{\boldsymbol{\Sigma}} &= (1-\hat{\rho})\mathbf{I}_p + \hat{\rho}\mathbf{1}\mathbf{1}^T\end{aligned}$$

   can be assumed to have the same set of eigenvectors. So, $\boldsymbol{\Sigma} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^T$ and $\hat{\boldsymbol{\Sigma}} = \mathbf{P}\hat{\boldsymbol{\Lambda}}\mathbf{P}^T$, where $\boldsymbol{\Lambda} = \text{Diag}(1 + \overline{p-1}\rho, 1-\rho, \ldots, 1-\rho)$ and

$\hat{\mathbf{\Lambda}} = \text{Diag}(1+\overline{p-1}\hat{\rho}, 1-\hat{\rho}, \ldots, 1-\hat{\rho})$. Hence, $\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1} = \mathbf{P}\mathbf{\Lambda}\hat{\mathbf{\Lambda}}^{-1}\mathbf{P}^T$, and $k = \lambda_{max}(\mathbf{R})/\lambda_{min}(\mathbf{R}) = \lambda_{max}(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1})/\lambda_{min}(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1}) = (1+\overline{p-1}\rho)(1+\overline{p-1}\hat{\rho})^{-1}(1-\rho)(1-\hat{\rho})^{-1}$ or $(1-\rho)(1-\hat{\rho})^{-1}(1+\overline{p-1}\rho)(1+\overline{p-1}\hat{\rho})^{-1}$. So, when ρ and $\hat{\rho}$ are near to each other, then r will not be very high.

3. One major problem of high dimensional classification problem is accurate estimation of small eigenvalues of $\mathbf{\Sigma}$. In the classical estimate of $\mathbf{\Sigma}$, these eigenvalues are underestimated. To cope with this problem, if we replace the small eigenvalues by a small number ϵ, that may lead to another regularization. To study this regularization, we now consider the best case scenario in this approach i.e., the case when all the eigenvalues except the small ones and all the eigenvectors have been correctly estimated. So, then $\mathbf{\Sigma} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$ and $\hat{\mathbf{\Sigma}} = \mathbf{P}\hat{\mathbf{\Lambda}}\mathbf{P}^T$, where, $\mathbf{\Lambda} = \text{Diag}(\lambda_1, \ldots, \lambda_t, \lambda_{t+1}, \ldots, \lambda_p)$ and $\hat{\mathbf{\Lambda}} = \text{Diag}(\lambda_1, \ldots, \lambda_t, \epsilon, \ldots, \epsilon)$. Hence, $\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1} = \mathbf{P}\mathbf{\Lambda}\hat{\mathbf{\Lambda}}^{-1}\mathbf{P}^T$, and $k = \lambda_{max}(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1})/\lambda_{min}(\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1}) = \frac{\max(1,\lambda_{t+1}/\epsilon,\ldots,\lambda_p/\epsilon)}{\min(1,\lambda_{t+1}/\epsilon,\ldots,\lambda_p/\epsilon)}$. Clearly, $k \geq \frac{\lambda_{t+1}/\epsilon}{\lambda_p/\epsilon} = \frac{\lambda_{t+1}}{\lambda_p}$. So, if the small eigenvalues maintain very high ratio between them, k can still be large for any choice of ϵ.

4. When $\mathbf{\Sigma}$ is block-diagonal matrix with blocks having intraclass correlation structures i.e.,

$$\mathbf{\Sigma} = \begin{bmatrix} (1-\gamma)\mathbf{I}_p + \gamma\mathbf{J}_p & \mathbf{0} \\ \mathbf{0} & (1-\gamma)\mathbf{I}_p + \gamma\mathbf{J}_p \end{bmatrix}$$

and we estimate it by

$$\hat{\mathbf{\Sigma}} = (1-\rho)\mathbf{I}_{2p} + \rho\mathbf{J}_{2p},$$

then the matrix $\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1}$ equals $\mathbf{\Sigma}(a\mathbf{I}_{2p} + b\mathbf{J}_{2p}$, where $a = 1/(1-\rho)$ and $b = -\rho/\{(1-\rho)(1+\overline{2p-1}\rho)\}$. So,

$$\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1} = a\mathbf{\Sigma} + b\mathbf{\Sigma}\mathbf{J}_{2p} = a(1-\gamma)\mathbf{I}_{2p} + a\gamma\mathbf{K} + b(1+\overline{p-1}\gamma)\mathbf{J}_{2p},$$

where

$$\mathbf{K} = \begin{bmatrix} \mathbf{J}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_p \end{bmatrix}.$$

So, the distinct eigenvalues of $\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1}$ are $a(1-\gamma), (1+\overline{p-1}\gamma)(a+2pb), a(1+\overline{p-1}\gamma)$, as the distinct eigenvalues of $c\mathbf{K} + d\mathbf{J}_{2p}$ are $0, p(c+2d), pc$. Putting the values of a and b, the eigenvalues of $\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-1}$ are $\frac{1-\gamma}{1-\rho}, \frac{1+\overline{p-1}\gamma}{1+\overline{2p-1}\rho}$ and $\frac{1+\overline{p-1}\gamma}{1-\rho}$. The first two are the ratio of comparable eigenvalues of $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}$, but the third one can be very high if p is large leading to a very high value of k.

3.2.3 Singular $\hat{\Sigma}$

When $\hat{\mathbf{\Sigma}}$, the estimate of $\mathbf{\Sigma}$, is singular, which is quite likely in case of high dimensional classification problems, the linear discriminant score is not well defined, because $\hat{\mathbf{\Sigma}}^{-1}$ is not available. In such situations, generally the Moore-Penrose g-inverse of $\hat{\mathbf{\Sigma}}$ is used (Ref. [Bickel & Levina, 2004]) in place of $\mathbf{\Sigma}^{-1}$ in the expression of the Bayes classifier. Then the corresponding misclassification probability is

$$\overline{\Phi}\left(\frac{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}^{-}\boldsymbol{\mu}}{2\sqrt{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}^{-}\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^{-}\boldsymbol{\mu}}}\right)$$

Since, $\hat{\mathbf{\Sigma}}$ is singular, $C\left(\hat{\mathbf{\Sigma}}\right)^{\perp} \neq \emptyset$ So, if $\boldsymbol{\mu} \in C\left(\hat{\mathbf{\Sigma}}^{-}\right)^{\perp}$, the Bayes classifier $\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}^{-}(\mathbf{X} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2) = 0$, which forces the misclassification probability to be 1. But $C\left(\hat{\mathbf{\Sigma}}\right)^{\perp}$ is a lower dimensional subspace of $\mathbf{R}^p$. Now, we will study the effect of using $\hat{\mathbf{\Sigma}}^{-}$ when $\boldsymbol{\mu} \notin C\left(\hat{\mathbf{\Sigma}}^{-}\right)^{\perp}$.

Let, $\hat{\mathbf{\Sigma}}^{-} = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A}$ is $p \times k (k < p)$ and has full column rank. As earlier, we will consider the ratio of t he arguments of $\overline{\Phi}$ in the expression of misclassification probability of the Bayes classifier and the linear classifier corresponding to $\hat{\mathbf{\Sigma}}^{-}$. The ratio is

$$\begin{aligned}
r =& \frac{\sqrt{\boldsymbol{\mu}^T\mathbf{A}\mathbf{A}^T\mathbf{\Sigma}A\mathbf{A}^T\boldsymbol{\mu}}\sqrt{\boldsymbol{\mu}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}}}{\boldsymbol{\mu}^T\mathbf{A}\mathbf{A}^T\boldsymbol{\mu}} \\
=& \frac{\sqrt{\mathbf{u}^T\mathbf{\Sigma}^{\frac{1}{2}}A\mathbf{A}^T\mathbf{\Sigma}A\mathbf{A}^T\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{u}}\sqrt{\mathbf{u}^T\mathbf{u}}}{\mathbf{u}^T\mathbf{\Sigma}^{\frac{1}{2}}A\mathbf{A}^T\mathbf{\Sigma}^{\frac{1}{2}}\mathbf{u}} && \text{(where } \mathbf{u} = \mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}\text{)} \\
=& \frac{\sqrt{\mathbf{u}^T\mathbf{B}\mathbf{B}^T\mathbf{B}\mathbf{B}^T\mathbf{u}}\sqrt{\mathbf{u}^T\mathbf{u}}}{\mathbf{u}^T\mathbf{B}\mathbf{B}^T\mathbf{u}} && \text{(where } \mathbf{B} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{A}\text{).}
\end{aligned}$$

Let us split $\mathbf{u}$ into two components $\mathbf{u}_1$ and $\mathbf{u}_2$ such that $\mathbf{u} = \mathbf{u}_1 \oplus \mathbf{u}_2$ and $\mathbf{u}_1 \in c(\mathbf{B}), \mathbf{u}_2 \in c(\mathbf{B})^{\perp}$. Then

$$
\begin{aligned}
r^2 =& \frac{\mathbf{u}_1^T\mathbf{B}\mathbf{B}^T\mathbf{B}\mathbf{B}^T\mathbf{u}_1.(\mathbf{u}_1^T\mathbf{u}_1 + \mathbf{u}_2^T\mathbf{u}_2)}{\mathbf{u}_1^T\mathbf{B}\mathbf{B}^T\mathbf{u}_1} && (\text{as, } \mathbf{u}_1 \perp \mathbf{u}_2, \mathbf{u}^T\mathbf{u} = \mathbf{u}_1^T\mathbf{u}_1 + \mathbf{u}_2^T\mathbf{u}_2) \\
=& \frac{\mathbf{u}_1^T\mathbf{B}\mathbf{B}^T\mathbf{B}\mathbf{B}^T\mathbf{u}_1.\mathbf{u}_1^T\mathbf{u}_1}{\mathbf{u}_1^T\mathbf{B}\mathbf{B}^T\mathbf{u}_1} \frac{\mathbf{u}_1^T\mathbf{u}_1 + \mathbf{u}_2^T\mathbf{u}_2}{\mathbf{u}_1^T\mathbf{u}_1} \\
=& \frac{\mathbf{v}^T(\mathbf{B}^T\mathbf{B})\mathbf{v}.\mathbf{v}^T(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{v}}{\mathbf{v}^T v} \frac{\mathbf{u}_1^T\mathbf{u}_1 + \mathbf{u}_2^T\mathbf{u}_2}{\mathbf{u}_1^T\mathbf{u}_1} && [\text{as } \mathbf{u}_1 \in C(\mathbf{B}), \text{ (say), } \mathbf{u}_1 = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{v}] \\
=& \frac{\mathbf{v}^T(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A})\mathbf{v}.\mathbf{v}^T(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{v}}{\mathbf{v}^T v} \frac{\mathbf{u}_1^T\mathbf{u}_1 + \mathbf{u}_2^T\mathbf{u}_2}{\mathbf{u}_1^T\mathbf{u}_1} && (\text{as } \mathbf{B} = \mathbf{\Sigma}^{\frac{1}{2}}\mathbf{A}) \\
\leq& K\frac{\mathbf{u}_1^T\mathbf{u}_1 + \mathbf{u}_2^T\mathbf{u}_2}{\mathbf{u}_1^T\mathbf{u}_1} && \left[\text{where} K = \frac{2\sqrt{k}}{1+k} \text{ and } k = \frac{\lambda_{max}(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A})}{\lambda_{min}(\mathbf{A}^T\mathbf{\Sigma}\mathbf{A})}\right].
\end{aligned}
$$

The upper bound of the first term is attainable, but the second term is unbounded. So r^2 can achieve any value in $(0, \infty)$. Hence when $\hat{\mathbf{\Sigma}}$ is singular, the misclassification probability can take any value up to .5.

Chapter 4

A New Method of Regularization

In the case of high dimensional data, generally the covariance matrix $\boldsymbol{\Sigma}$ is regularized by imposing some restrictions on it (e.g., by ignoring all correlations or assuming all correlations to be equal). A more flexible form of regularization is obtained by considering a convex combination of sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ and the MLE of $\boldsymbol{\Sigma}$ under the restrictions mentioned. However, in the expression of the linear classifier, $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2)$, the contribution of $\boldsymbol{\Sigma}$ is through $\boldsymbol{\Sigma}^{-1}$. So, a reasonable choice of regularization is to consider a convex combination of $\hat{\boldsymbol{\Sigma}}^{-1}$ (or $\hat{\boldsymbol{\Sigma}}^{-}$ if $\hat{\boldsymbol{\Sigma}}$ is singular) and the inverse of the MLE of $\boldsymbol{\Sigma}$ under different restrictions. In this chapter, we will study the performance of this kind of regularization. The choice of the generalized inverse should not matter as it is used only in a quadratic form that is invariant under this choice. In order to avoid confusion, we use $\hat{\boldsymbol{\Sigma}}^{+}$, the Moore-Penrose generalized inverse in the rest of this dissertation.

4.1 Optimal Linear Combination of Two $\boldsymbol{\Sigma}^{-1}$-estimates

Let us consider two estimators of $\boldsymbol{\Sigma}$, namely $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$, and we will consider a linear combination of $\hat{\boldsymbol{\Sigma}}_1^{+}$ and $\hat{\boldsymbol{\Sigma}}_2^{+}$, $\lambda_1\hat{\boldsymbol{\Sigma}}_1^{+} + \lambda_2\hat{\boldsymbol{\Sigma}}_2^{+}$ in place of $\boldsymbol{\Sigma}^{-1}$ in the expression of the linear classifier. Since, we are concentrating on regularization of $\boldsymbol{\Sigma}$, we will assume that the mean separation vector $\boldsymbol{\mu}$ to be known. We define $S(\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$ to be the set of all linear classifier with $\boldsymbol{\Sigma}^{-1}$ replaced by $\lambda_1\hat{\boldsymbol{\Sigma}}_1^{+} + \lambda_2\hat{\boldsymbol{\Sigma}}_2^{+}$ for $\lambda_1, \lambda_2 \in \mathbb{R}$

Proposition 4.1.1. *For the classification problem with two populations and known mean-separation vector, for any two estimators $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ of $\mathbf{\Sigma}$, the misclassification probability of the optimal classifier in the class $S(\hat{\mathbf{\Sigma}}_1, \hat{\mathbf{\Sigma}}_2)$ is given by*

$$\overline{\Phi}\left(\frac{\sqrt{\|\mathbf{P}_{\mathbf{Y}}(\mathbf{x})\|^2}}{2}\right), \tag{4.1.1}$$

where $\mathbf{x} = \mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}, \mathbf{Y} = \begin{bmatrix}\mathbf{y} & \mathbf{z}\end{bmatrix}$ for $\mathbf{y} = \mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_1^{+}\boldsymbol{\mu}, \mathbf{z} = \mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_2^{+}\boldsymbol{\mu}$. Here, $\mathbf{P}_{\mathbf{Y}}$ is the projection operator into the column-space of $\mathbf{Y}$.

Proof. Misclassification probability for using $\hat{\mathbf{\Sigma}}_1^{+}$ and $\hat{\mathbf{\Sigma}}_2^{+}$ are $\overline{\Phi}\left[\frac{a(\hat{\mathbf{\Sigma}}_1)}{2}\right]$ and $\overline{\Phi}\left[\frac{a(\hat{\mathbf{\Sigma}}_2)}{2}\right]$ respectively, where

$$a(\hat{\mathbf{\Sigma}}_1) = \frac{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_1^{+}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_1^{+}\mathbf{\Sigma}\hat{\mathbf{\Sigma}}_1^{+}\boldsymbol{\mu}}} = \frac{(\boldsymbol{\mu}^T\mathbf{\Sigma}^{-\frac{1}{2}})(\mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_1^{+}\boldsymbol{\mu})}{\sqrt{(\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_1^{+}\mathbf{\Sigma}^{\frac{1}{2}})(\mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_1^{+}\boldsymbol{\mu})}} = \frac{\mathbf{x}^T\mathbf{y}}{\sqrt{\mathbf{y}^T\mathbf{y}}}, \tag{4.1.2}$$

and

$$a(\hat{\mathbf{\Sigma}}_2) = \frac{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_2^{+}\boldsymbol{\mu}}{\sqrt{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_2^{+}\mathbf{\Sigma}\hat{\mathbf{\Sigma}}_2^{+}\boldsymbol{\mu}}} = \frac{(\boldsymbol{\mu}^T\mathbf{\Sigma}^{-\frac{1}{2}})(\mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_2^{+}\boldsymbol{\mu})}{\sqrt{(\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_2^{+}\mathbf{\Sigma}^{\frac{1}{2}})(\mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_2^{+}\boldsymbol{\mu})}} = \frac{\mathbf{x}^T\mathbf{z}}{\sqrt{\mathbf{z}^T\mathbf{z}}}, \tag{4.1.3}$$

and $\mathbf{x} = \mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}, \mathbf{y} = \mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_1^{+}\boldsymbol{\mu}, \mathbf{z} = \mathbf{\Sigma}^{\frac{1}{2}}\hat{\mathbf{\Sigma}}_2^{+}\boldsymbol{\mu}$. When $\lambda_1\hat{\mathbf{\Sigma}}_1^{+} + \lambda_2\hat{\mathbf{\Sigma}}_2^{+}$ is used as an estimate of $\mathbf{\Sigma}^{-1}$, the resulting misclassification probability will be $\overline{\Phi}\left[\frac{f(\lambda_1,\lambda_2)}{2}\right]$, where

$$f(\lambda_1, \lambda_2) = \frac{\mathbf{x}^T(\lambda_1\mathbf{y} + \lambda_2\mathbf{z})}{\sqrt{(\lambda_1\mathbf{y} + \lambda_2\mathbf{z})^T(\lambda_1\mathbf{y} + \lambda_2\mathbf{z})}}. \tag{4.1.4}$$

So,

$$\begin{aligned} f^2(\lambda_1, \lambda_2) =& \frac{(\boldsymbol{\lambda}^T\mathbf{d})^2}{\boldsymbol{\lambda}^T\mathbf{A}\boldsymbol{\lambda}} \quad \left[\text{where } \boldsymbol{\lambda} = \begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}, \mathbf{d} = \begin{pmatrix}\mathbf{x}^T\mathbf{y}\\ \mathbf{x}^T\mathbf{z}\end{pmatrix} \text{and } \mathbf{A} = \begin{pmatrix}\mathbf{y}^T\mathbf{y} & \mathbf{y}^T\mathbf{z}\\ \mathbf{y}^T\mathbf{z} & \mathbf{z}^T\mathbf{z}\end{pmatrix}\right] \\ \leq& \mathbf{d}^T\mathbf{A}^{-1}\mathbf{d} \\ =& \mathbf{x}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{x} \qquad \left[\text{as } \mathbf{d} = \mathbf{Y}^T\mathbf{x}, \mathbf{A} = \mathbf{Y}^T\mathbf{Y}\right] \\ =& \|\mathbf{P}_{\mathbf{Y}}(\mathbf{x})\|^2. \end{aligned} \tag{4.1.5}$$

Equality holds when $\boldsymbol{\lambda} \propto \mathbf{A}^{-1}\mathbf{d}$. Hence, the optimal classifier in the class $S(\hat{\mathbf{\Sigma}}_1, \hat{\mathbf{\Sigma}}_2)$ has misclassification probability $\overline{\Phi}\left(\frac{\sqrt{\|\mathbf{P}_{\mathbf{Y}}(\mathbf{x})\|^2}}{2}\right)$. □

But the optimal $\boldsymbol{\lambda}$ may not correspond to a convex combination of $\hat{\boldsymbol{\Sigma}}_1^+$ and $\hat{\boldsymbol{\Sigma}}_2^+$ unless both the components of $\mathbf{A}^{-1}\mathbf{d}$ have the same sign. However, for this optimal $\boldsymbol{\lambda}$, we get a linear classifier, where the matrix multiplying $\mathbf{x}$ is possibly not n.n.d, with misclassification probability smaller than that of both $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$.

4.2 Optimal Convex Combination of Two $\boldsymbol{\Sigma}^{-1}$-estimates

It will be interesting to know the condition which makes the optimal linear combination a convex combination. The next proposition gives a necessary and sufficient condition for a convex combination to be an optimal classifier.

Proposition 4.2.1. *The necessary and sufficient condition for a proper convex combination of $\hat{\boldsymbol{\Sigma}}_1^+$ and $\hat{\boldsymbol{\Sigma}}_2^+$ to be an optimal classifier in the class $S(\hat{\boldsymbol{\Sigma}}_1^+, \hat{\boldsymbol{\Sigma}}_2^+)$ is*

$$\frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}} < \frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}} < \frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}} \tag{4.2.1}$$

Moreover, violation of the right-inequality (or left-inequality) will force $\hat{\boldsymbol{\Sigma}}_1^+$ (or $\hat{\boldsymbol{\Sigma}}_2^+$) to be the best among the convex combinations os $\hat{\boldsymbol{\Sigma}}_1^-$ and $\hat{\boldsymbol{\Sigma}}_2^-$

Proof. We observe that if we consider the change of variable $\frac{\lambda_1}{\lambda_2} = u$, then

$$\begin{aligned} f^2(\lambda_1, \lambda_2) =& \frac{\lambda_1^2 d_1^2 + 2\lambda_1\lambda_2 d_1 d_2 + \lambda_2^2 d_2^2}{\lambda_1^2 a_{11} + 2\lambda_1\lambda_2 a_{12} + \lambda_2^2 a_{22}} \qquad (4.2.2) \\ & \qquad [\text{where } \mathbf{d} = (d_1, d_2)^T \text{ and } \mathbf{A} = ((a_{ij}))] \\ =& \frac{u^2 d_1^2 + 2u d_1 d_2 + d_2^2}{u^2 a_{11} + 2u a_{12} + a_{22}} \\ =& \frac{g(u)}{h(u)} \qquad \text{(say)}. \qquad (4.2.3) \end{aligned}$$

The optimal λ will correspond to a convex combination if the maximum of $\frac{g(u)}{h(u)}$ is nonnegative. To find the maximum of $f^2(\lambda_1, \lambda_2)$, we put $\frac{\partial}{\partial u}\left[\frac{g(u)}{h(u)}\right] = 0$.

Now,

$$\begin{aligned}\frac{\partial}{\partial u}\left[\frac{g(u)}{h(u)}\right] =& \frac{h(u)g^{'}(u) - g(u)h^{'}(u)}{h^2(u)} \\ =& \frac{[a_{12}d_1^2 - a_{11}d_1d_2]u^2 + [a_{22}d_1^2 - a_{11}d_2^2]u + [a_{22}d_1d_2 - a_{12}d_2^2]}{h^2(u)} \\ =& \frac{q(u)}{h^2(u)} \qquad \text{(say).} \end{aligned} \tag{4.2.4}$$

So, we need to find the zeros of the quadratic equation $q(u) = 0$. The discriminant of this quadratic equation is

$$\begin{aligned}&[a_{22}d_1^2 - a_{11}d_2^2]^2 - 4[a_{12}d_1^2 - a_{11}d_1d_2][a_{22}d_1d_2 - a_{12}d_2^2] \\ =&[a_{22}d_1^2 - a_{11}d_2^2]^2 + 4[a_{11}a_{22} + a_{12}^2]d_1^2d_2^2 - 4a_{12}d_1d_2[a_{22}d_1^2 + a_{11}d_2^2] \\ \geq&[a_{22}d_1^2 - a_{11}d_2^2]^2 + 4[2\sqrt{a_{11}a_{22}}a_{12}]d_1^2d_2^2 - 4a_{12}d_1d_2[a_{22}d_1^2 + a_{11}d_2^2] \\ &\qquad \text{(By A.M-G.M inequality } a_{11}a_{22} + a_{12}^2 \geq 2\sqrt{a_{11}a_{22}}|a_{12}| \text{ and } |a_{12}| \geq a_{12}) \\ =&[a_{22}d_1^2 - a_{11}d_2^2]^2 - 4a_{12}d_1d_2[\sqrt{a_{22}}d_1 - \sqrt{a_{11}}d_2]^2 \\ =&[a_{22}d_1^2 - a_{11}d_2^2]^2 - 4\sqrt{a_{11}a_{22}}d_1d_2[\sqrt{a_{22}}d_1 - \sqrt{a_{11}}d_2]^2 \quad (\text{as } a_{12} \leq |a_{12}| \leq \sqrt{a_{11}a_{22}}) \\ =&[\sqrt{a_{22}}d_1 - \sqrt{a_{11}}d_2]^2([\sqrt{a_{22}}d_1 + \sqrt{a_{11}}d_2]^2 - 4\sqrt{a_{11}a_{22}}d_1d_2) \\ =&[\sqrt{a_{22}}d_1 - \sqrt{a_{11}}d_2]^4 \\ \geq&0 \end{aligned} \tag{4.2.5}$$

This derivation shows that the discriminant of $q(u)$ is always positive unless $|A| = 0$ and $\frac{d_1}{\sqrt{a_{11}}} = \frac{d_2}{\sqrt{a_{22}}}$; which is a very improbable condition. So, we assume that the discriminant is positive and hence $q(u)$ has two distinct roots, which correspond to the maximum and the minimum of $f^2(\lambda_1, \lambda_2)$. The optimal λ_1, λ_2 will have the same sign, when the root of $q(u)$ corresponding to the maximum of $f^2(\lambda_1, \lambda_2)$ is nonnegative.

Next, we observe that if the coefficient of u^2 in $q(u)$ is nonnegative, i.e.,

$$\begin{aligned}&a_{12}d_1^2 - a_{11}d_1d_2 \geq 0 \\ \Rightarrow&a_{12} \geq a_{11}\frac{d_2}{d_1} \qquad (\text{as } d_1, d_2 > 0) \\ \Rightarrow&a_{22}\frac{d_1}{d_2} > a_{12} \quad (\text{as otherwise } a_{11}\frac{d_2}{d_1}, a_{22}\frac{d_1}{d_2} \leq a_{12} \Rightarrow a_{11}a_{22} \leq a_{12}^2, \text{ which is false}) \\ \Rightarrow&a_{22}\frac{d_1}{d_2} > a_{12} \geq a_{11}\frac{d_2}{d_1} \\ \Rightarrow&a_{22}d_1^2 - a_{11}d_2^2 > 0 \text{ and } a_{22}d_1d_2 - a_{12}d_2^2 > 0. \end{aligned} \tag{4.2.6}$$

Then, all coefficients in $q(u)$ are positive, which implies that both the roots of $q(u)$ are nonpositive and $q(u) > 0$ on (o, ∞). So, $f^2(\lambda_1, \lambda_2)$ (under the restriction λ_1, λ_2 have the same sign) attains its maximum at $\lambda_2 = 0$ (i.e., $u = \infty$). So, in this case $\hat{\boldsymbol{\Sigma}}_1^+$ is the best among all the convex combinations of $\hat{\boldsymbol{\Sigma}}_1^+$ and $\hat{\boldsymbol{\Sigma}}_2^+$. Similarly, when the coefficient of u^2 in $q(u)$ is negative, i.e., $a_{12} < a_{11}\frac{d_2}{d_1}$, and also the constant term is nonpositive, i.e., $a_{12} \leq a_{11}\frac{d_2}{d_1}$ then all coefficients of $q(u)$ are negative, which means that both the roots of $q(u)$ are nonpositive and $q(u) < 0$ on $(0, \infty)$. So, $f^2(\lambda_1, \lambda_2)$ (under the same restriction) attains the maximum at $\lambda_1 = 0$ (i.e. $u = 0$). Then, $\hat{\boldsymbol{\Sigma}}_2^+$ is the best among all convex combinations of $\hat{\boldsymbol{\Sigma}}_1^+$ and $\hat{\boldsymbol{\Sigma}}_2^+$. Excluding these two cases, $f^2(\lambda_1, \lambda_2)$ will attain its maximum for some λ_1, λ_2 having the same sign (i.e., some matrix, which is a strict convex combination of $\hat{\boldsymbol{\Sigma}}_1^+$ and $\hat{\boldsymbol{\Sigma}}_2^+$, will be the best) because larger root of $q(u)$, which corresponds to the maximum of $\frac{g(u)}{h(u)}$, is positive. Hence, a necessary and sufficient condition for a strict convex combination of $\hat{\boldsymbol{\Sigma}}_1^+$ and $\hat{\boldsymbol{\Sigma}}_2^+$ to be the best, us the following

$$
\begin{aligned}
& a_{12}d_1^2 - a_{11}d_1d_2 < 0 \ \text{ and } \ a_{22}d_1d_2 > a_{12}d_2^2 \\
\Leftrightarrow \quad & a_{12} < a_{11}\frac{d_2}{d_1}, a_{22}\frac{d_1}{d_2} \\
\Leftrightarrow \quad & \frac{a_{12}}{a_{22}} < \frac{d_1}{d_2} < \frac{a_{11}}{a_{12}} \\
\Leftrightarrow \quad & \frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}} < \frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}} < \frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}_2^+\boldsymbol{\mu}}.
\end{aligned} \tag{4.2.7}
$$

If the right-inequality is violated, then $a_{12}d_1 \geq a_{11}d_2$, which implies the coefficient of u^2 in $q(u)$ is nonnegative, which forces $\hat{\boldsymbol{\Sigma}}_1^+$ to be the best (as shown earlier). Similarly, if the left-inequality is violated, then the coefficient of u^2 and the constant term in $q(u)$ are nonpositive, and hence $\hat{\boldsymbol{\Sigma}}_2^+$ is forced to be the best (as shown earlier). As $\frac{d_1}{d_2}$ goes away from $\frac{a_{11}}{a_{12}}$ towards $\frac{a_{12}}{a_{22}}$, the optimal classifier in the class $S(\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2)$ moves away from that corresponding to $\hat{\boldsymbol{\Sigma}}_1^+$ towards the classifier corresponding to $\hat{\boldsymbol{\Sigma}}_2^+$ and vice versa. □

In the next subsection we give illustrations which favor moving away from sample covariance matrix towards the corresponding diagonal or intra-class correlation matrix in our shrinkage method, considered in this section.

4.3 Illustrations Favoring Shrinkage Methods

In the case of high dimensional classification problems, the difficulty lies mostly in estimating the correlation matrix properly. So, if the true $\mathbf{\Sigma} = \mathbf{DRD}$, where $\mathbf{R}$ is the true correlation matrix and $\mathbf{D}$ is a diagonal matrix consisting of the standard deviations, and the sample based covariance matrix is $\hat{\mathbf{\Sigma}} = \hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}}$, where $\hat{\mathbf{R}}$ and $\hat{\mathbf{D}}$ are the corresponding estimates, in most of the high dimensional situations $\hat{\mathbf{D}}$ is reasonably close to $\mathbf{D}$, But $\hat{\mathbf{R}}$ is away from $\mathbf{R}$ (both in terms of eigenvalues and eigenvectors). Also, sometimes $\hat{\mathbf{R}}$ has rank deficiency. The best possible scenario will be $\hat{\mathbf{D}} = \mathbf{D}$ and $\hat{\mathbf{R}} = \mathbf{R}_+$, where $\mathbf{R}_+$ is a lower rank approximation of $\mathbf{R}$ (i.e $\mathbf{R}_+$ has all eigenvalues and eigenvectors same as those of $\mathbf{R}$ except few small eigenvalues, which are zeros for $\mathbf{R}_+$ but positive for $\mathbf{R}$). In this section, we will consider this special case and study the behavior of the discussed shrinkage estimates of dispersion matrix and the associated classifiers.

4.3.1 Shrinkage Towards Diagonal Matrix

In this section, we will consider the shrinkage towards $\text{Diag}(\hat{\mathbf{\Sigma}})$, i.e., a diagonal matrix with diagonal entries same as those of $\hat{\mathbf{\Sigma}}$. If the spectral decomposition of $\mathbf{R}$ is $\mathbf{R} = \mathbf{P\Lambda P}^T$, then $\hat{\mathbf{R}} = \mathbf{R}_+ = \mathbf{P\Lambda}_+\mathbf{P}^T$, where $\mathbf{\Lambda} = \text{Diag}(\lambda_1, \ldots, \lambda_p)$, and $\mathbf{\Lambda}_+ = \text{Diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)$ for some $1 \leq k \leq p-1$. In this situation,

$$
\begin{aligned}
&\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}^-\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^-\boldsymbol{\mu} \\
=&\boldsymbol{\mu}^T\mathbf{D}^{-1}\mathbf{R}_+^-\mathbf{D}^{-1}\mathbf{DRDD}^{-1}\mathbf{R}_+^-\mathbf{D}^{-1}\boldsymbol{\mu} \\
=&\boldsymbol{\mu}^T(\mathbf{D}^{-1}\mathbf{P\Lambda}_+^-\mathbf{P}^T\mathbf{D}^{-1})(\mathbf{DP\Lambda P}^T\mathbf{D})(\mathbf{D}^{-1}\mathbf{P\Lambda}_+^-\mathbf{P}^T\mathbf{D}^{-1})\boldsymbol{\mu} \\
=&\boldsymbol{\nu}^T\mathbf{\Lambda}_+^-\mathbf{\Lambda\Lambda}_+^-\boldsymbol{\nu} \qquad \text{(putting } \boldsymbol{\nu} = \mathbf{P}^T\mathbf{D}^{-1}\boldsymbol{\mu} \text{ and using } \mathbf{P}^T\mathbf{P} = \mathbf{I}_p\text{)} \\
=&\boldsymbol{\nu}^T\mathbf{\Lambda}_+^-\boldsymbol{\nu} \\
=&\sum_{i=1}^{k}\frac{1}{\lambda_i}\nu_i^2.
\end{aligned}
$$

Using similar arguments, we have

$$\begin{aligned}
\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}^{-}\boldsymbol{\mu} = \boldsymbol{\nu}^T\boldsymbol{\Lambda}_{+}^{-}\boldsymbol{\nu} &= \sum_{i=1}^{k}\frac{1}{\lambda_i}\nu_i^2 \\
\boldsymbol{\mu}^T\hat{\mathbf{D}}^{-2}\boldsymbol{\mu} \qquad &= \boldsymbol{\nu}^T\boldsymbol{\nu} = \sum_{i=1}^{p}\nu_i^2 \\
\boldsymbol{\mu}^T\hat{\mathbf{D}}^{-2}\boldsymbol{\Sigma}\hat{\mathbf{D}}^{-2}\boldsymbol{\mu} \qquad &= \boldsymbol{\nu}^T\boldsymbol{\Lambda}\boldsymbol{\nu} = \sum_{i=1}^{p}\lambda_i\nu_i^2 \\
\boldsymbol{\mu}^T\hat{\mathbf{D}}^{-2}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-}\boldsymbol{\mu} = \boldsymbol{\nu}^T\boldsymbol{\Lambda}\boldsymbol{\Lambda}_{+}^{-}\boldsymbol{\nu} &= \sum_{i=1}^{k}\nu_i^2
\end{aligned}$$

Now we apply Proposition 4.2.1 with $\hat{\boldsymbol{\Sigma}}_1 = \hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}_2 = \hat{\mathbf{D}}^2$. We observe that

$$\frac{\sum_{i=1}^{k}\frac{1}{\lambda_i}\nu_i^2}{\sum_{i=1}^{p}\nu_i^2} = \frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}^{-}\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\mathbf{D}}^{-2}\boldsymbol{\mu}} = \frac{d_1}{d_2} < \frac{a_{11}}{a_{12}} = \frac{\boldsymbol{\mu}^T\hat{\boldsymbol{\Sigma}}^{-}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-}\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\mathbf{D}}^{-2}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-}\boldsymbol{\mu}} = \frac{\sum_{i=1}^{k}\frac{1}{\lambda_i}\nu_i^2}{\sum_{i=1}^{k}\nu_i^2}. \quad (4.3.1)$$

So, even if $k = p - 1$ (i.e., all eigenvectors and eigenvalues of the true correlation matrix $\mathbf{R}$ except only the smallest eigenvalue are known), it is better to use the discussed shrinkage method towards diagonal matrix, when $\boldsymbol{\Sigma}$ is nonsingular and the variances are perfectly estimated.

Next, we notice

$$\frac{a_{12}}{a_{22}} < \frac{d_1}{d_2} \quad (4.3.2)$$

$$\Leftrightarrow \frac{\sum_{i=1}^{k}\nu_i^2}{\sum_{i=1}^{p}\lambda_i\nu_i^2} < \frac{\sum_{i=1}^{k}\frac{1}{\lambda_i}\nu_i^2}{\sum_{i=1}^{p}\nu_i^2} \quad (4.3.3)$$

$$\Leftrightarrow \left(\sum_{i=1}^{k}\nu_i^2\right)\left(\sum_{i=1}^{p}\nu_i^2\right) < \left(\sum_{i=1}^{k}\frac{1}{\lambda_i}\nu_i^2\right)\left(\sum_{i=1}^{p}\lambda_i\nu_i^2\right) \quad (4.3.4)$$

$$\Leftrightarrow 0 < \sum_{1\le i<j\le k}\nu_i^2\nu_j^2\left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} - 2\right) + \sum_{1\le i\le k<j\le p}\nu_i^2\nu_j^2\left(\frac{\lambda_j}{\lambda_i} - 1\right) = S_1 + S_2 \quad \text{(say)}. \quad (4.3.5)$$

However, S_1 is always nonnegative and the second term is always nonpositive. So, nothing can be said about the above inequality in general. But if we

impose some restrictions on ν_is, the inequality will hold. For example, if ν_is can be taken to be equal (say, $\nu_i \approx \nu\ \forall i$), which means that the separation of means in different component are approximately the same after suitable scaling, the above inequality will hold in most of the cases because in that case

$$\begin{aligned}
&\sum_{1\le i<j\le k} \nu_i^2\nu_j^2\left(\frac{\lambda_i}{\lambda_j}+\frac{\lambda_j}{\lambda_i}-2\right)+\sum_{1\le i\le k<j\le p} \nu_i^2\nu_j^2\left(\frac{\lambda_j}{\lambda_i}-1\right)\\
&=\nu^4\left[\sum_{1\le i<j\le k}\left(\frac{\lambda_i}{\lambda_j}+\frac{\lambda_j}{\lambda_i}-2\right)+\sum_{1\le i\le k<j\le p}\left(\frac{\lambda_j}{\lambda_i}-1\right)\right]\\
&\ge\nu^4\left[\sum_{1\le i<j\le k}\left(\frac{\lambda_i}{\lambda_j}+\frac{\lambda_j}{\lambda_i}-2\right)-k(p-k)\right] \qquad (\text{as } 0\le\frac{\lambda_j}{\lambda_i}\forall i,j),
\end{aligned}$$

and all the summands of S_1 can be arbitrarily large as the ratio between the λ_i's increases. So, in most of the cases. the S_1 will be positive, which implies in these situations, the optimal linear combination of $\hat{\mathbf{\Sigma}}^+$ and $\hat{\mathbf{D}}^{-2}$ will correspond to a proper convex combination of them.

In another special case, which is rather pathological, if $\nu_i \approx 0, 1 \le i \le k$, then the inequality in (4.3.5) will not hold and hence the optimal convex combination will be the diagonal matrix itself. So, in general the convex combination is either the diagonal matrix itself or some proper convex combination of $\hat{\mathbf{\Sigma}}^+$ and $\mathbf{D}^{-2}$.

4.3.2 Shrinkage Towards An Intra-class Correlation Matrix

Here, we describe a situation where the shrinkage towards an intra-class correlation matrix is more advantageous than using the usual $\hat{\mathbf{\Sigma}}$. Let $\hat{\mathbf{\Sigma}}_C = \hat{\mathbf{D}}\hat{C}\hat{\mathbf{D}}$ be an estimate of $\mathbf{\Sigma}$ under the restriction that all correlations of $\mathbf{\Sigma}$ are equal, where $\hat{\mathbf{C}} = (1-\hat{\rho})\mathbf{I}_p + \hat{\rho}\mathbf{J}_p$. Under the special case discussed above, $\hat{\mathbf{D}} = \mathbf{D}$ and $\hat{\mathbf{\Sigma}} = \mathbf{D}\mathbf{R}_+\mathbf{D}$. Next, we observe that

$$\begin{aligned}
&\frac{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}^-\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_C^{-1}\boldsymbol{\mu}} < \frac{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}^-\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^-\boldsymbol{\mu}}{\boldsymbol{\mu}^T\hat{\mathbf{\Sigma}}_C^{-1}\mathbf{\Sigma}\hat{\mathbf{\Sigma}}^-\boldsymbol{\mu}}\\
&\Leftrightarrow \frac{\boldsymbol{\mu}^T\mathbf{D}^{-1}\mathbf{R}_+^-\mathbf{D}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}^T\mathbf{D}^{-1}\hat{\mathbf{C}}^{-1}\mathbf{D}^{-1}\boldsymbol{\mu}} < \frac{\boldsymbol{\mu}^T\mathbf{D}^{-1}\mathbf{R}_+^-\mathbf{R}\mathbf{R}_+^-\mathbf{D}^{-1}\boldsymbol{\mu}}{\boldsymbol{\mu}^T\mathbf{D}^{-1}\hat{\mathbf{C}}^{-1}\mathbf{R}\mathbf{R}_+^-\mathbf{D}^{-1}\boldsymbol{\mu}}\\
&\Leftrightarrow \boldsymbol{\mu}^T\mathbf{D}^{-1}\hat{\mathbf{C}}^{-1}\mathbf{D}^{-1}\boldsymbol{\mu} > \boldsymbol{\mu}^T\mathbf{D}^{-1}\hat{\mathbf{C}}^{-1}\mathbf{R}\mathbf{R}_+^-\mathbf{D}^{-1}\boldsymbol{\mu}\\
&\qquad (\text{as } \mathbf{R}_+^-\mathbf{R}\mathbf{R}_+^- = \mathbf{P}\mathbf{\Lambda}_+^-\mathbf{\Lambda}\mathbf{\Lambda}_+^-\mathbf{P}^T = \mathbf{P}\mathbf{\Lambda}_+^-\mathbf{P}^T = \mathbf{R}_+^-). \qquad (4.3.6)
\end{aligned}$$

Now, if all components of $\mathbf{D}^{-1}\boldsymbol{\mu}$ are approximately equal, which means that the standardized means are same, then $\hat{\mathbf{C}}^{-1}\mathbf{D}^{-1}\boldsymbol{\mu} = \lambda\mathbf{D}^{-1}\boldsymbol{\mu}$, and hence the inequality (4.3.6) holds, as

$$\boldsymbol{\mu}^T\mathbf{D}^{-1}\hat{\mathbf{C}}^{-1}\mathbf{D}^{-1}\boldsymbol{\mu} = \lambda\boldsymbol{\mu}^T\mathbf{D}^{-2}\boldsymbol{\mu} = \lambda\boldsymbol{\mu}^T\mathbf{D}^{-1}\mathbf{P}\mathbf{P}^T\mathbf{D}^{-1}\boldsymbol{\mu} = \lambda\sum_{i=1}^{p}\nu_i^2 \quad (\text{where, } \boldsymbol{\nu} = \mathbf{P}^T\mathbf{D}^{-1}\boldsymbol{\mu})$$

$$> \lambda\sum_{i=1}^{k}\nu_i^2 = \lambda.\boldsymbol{\nu}^T\boldsymbol{\Lambda}\boldsymbol{\Lambda}_+^-\boldsymbol{\nu} \qquad \left(\text{as,}\boldsymbol{\Lambda}\boldsymbol{\Lambda}_+^- = \begin{bmatrix}\mathbf{I}_k & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{bmatrix}\right)$$

$$= \lambda\boldsymbol{\mu}^T\mathbf{D}^{-1}\mathbf{R}\mathbf{R}_+^-\mathbf{D}^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}^T\mathbf{D}^{-1}\hat{\mathbf{C}}^{-1}\mathbf{R}\mathbf{R}_+^-\boldsymbol{\mu}.$$

So, by Proposition 4.2.1, shrinkage towards $\hat{\boldsymbol{\Sigma}}_C^{-1}$ away from $\hat{\boldsymbol{\Sigma}}^+$ is advantageous.

Thus we have obtained some small sample results in favor of regularization towards diagonal matrix and intra-class correlation matrix. The new shrinkage method, namely convex combination of the classical estimator of $\boldsymbol{\Sigma}^{-1}$ and an estimator with diagonal/intra-class correlation constraint, is quite promising, as it can reduce the misclassification probability considerably. A similar result is expected to hold for the convex combination of two estimates of $\boldsymbol{\Sigma}$ because both the shrinkage methods (namely, using convex combinations of $\hat{\boldsymbol{\Sigma}}_1^+, \hat{\boldsymbol{\Sigma}}_2^+$ in place of $\boldsymbol{\Sigma}^{-1}$, and using convex combinations of $\hat{\boldsymbol{\Sigma}}_1, \hat{\boldsymbol{\Sigma}}_2$ in place of $\boldsymbol{\Sigma}$) have the same extremes. So, if one kind of regularization is better than using the classical estimator, the other type of regularization should also be a reasonable strategy.

Chapter 5

Optimization of Variables for Clustering

As mentioned in section 1.3.2, the clustering problems become difficult, when the dimension p is large compared to the number of observations. There are mainly two kinds of problems which may arise because of high dimensionality: (a) effect of redundant dimensions when the true cluster structure remains confined to a much lower dimensional subspace, (b) statistical and computational challenges when data are insufficient relative to dimension. The first problem has been discussed in this chapter, the second will be addressed in Chapter 6.

As the simulation study described in the following section indicates, inclusion of too many noisy variables in the clustering procedure may produce clusters which differ from the actual ones substantially. For example, all the hierarchical clustering methods, which use Euclidean distance as the measure of dissimilarity, may perform badly in some case. We give a theoretical explanation of this phenomenon in Section 5.3.2 by showing that the presence of nondiscriminating variables weakens the distance matrix of the observations in a certain sense, thereby increasing the chances of wrong clusters. In such situations, dimension reduction strategy may be quite helpful in order to reduce the error. We give a framework for choosing suitable linear combinations of variables in model based clustering. In the first stage, we show how one can choose best linear combinations for a given number of such combinations. Then, we develop criteria for determining optimal number of linear combinations.

5.1 Difficulty of Clustering in Presence of Noisy Variables

In case of high dimensional data, if most of the variables are noisy, i.e., less relevant in the clustering context, and the true cluster structures are present only in a few variables, the hierarchical clustering methods may not be able to identify the clusters. For example, if single linkage or average linkage hierarchical method is used in such situations, it becomes difficult to identify the groups, even if the corresponding populations are substantially different from each other in terms of Mahalanobis distance. This phenomenon is reflected in the following simulation study.

In this study, three normal populations, each of size $n = 20$ and dimension $p = 60$, were considered. The common covariance matrix was the identity matrix. Separation of means was kept only in the first three components and in each of the remaining components, the same mean was used for all the populations. To observe the effect of noisy variables only, mean vectors of the three populations were suitably chosen to ensure very large Mahalanobis distance between the populations. Then average linkage hierarchical clustering method was used to identify the clusters. The experiment was repeated with $p = 20$ (including three dimensions where the separation of the population means was confined) and moderately large Mahalanobis distance between populations.

Results obtained from the simulation study are shown in the Table 5.1.1. The clusters mentioned in this table are obtained by applying a threshold on the hierarchical cluster tree that corresponds to exactly three clusters. The number reported in a 'cluster' column and a 'population' row is the number of elements of the cluster coming from that population.
As we learn from the simulation results, presence of too many noisy variables can create a great problem in recovering the clusters, even if the populations are sufficiently separated. So, in such situations we need to exclude the noisy variables and determine the discriminating variables (or more generally the discriminating directions) before applying any clustering method, in order to reduce the error.

5.2 Choosing Best Discriminating Linear Combinations of Variables

Generally, in order to select a certain number of discriminating linear combinations, a suitable criterion is considered. Among the possible candidates,

Mahalanobis Distance			Dimension							
			60				20			
P_1	P_2	P_3		$Clust_1$	$Clust_2$	$Clust_3$		$Clust_1$	$Clust_2$	$clust_3$
0			$Popul_1$	20	0	0	$Popul_1$	19	1	0
6	0		$Popul_2$	19	1	0	$Popul_2$	20	0	0
6	7	0	$Popul_3$	19	0	1	$Popul_3$	19	0	1
P_1	P_2	P_3		$Clust_1$	$Clust_2$	$Clust_3$		$Clust_1$	$Clust_2$	$Clust_3$
0			$Popul_1$	19	1	0	$Popul_1$	20	0	0
10	0		$Popul_2$	20	0	0	$Popul_2$	0	20	0
10	11.5	0	$Popul_3$	0	0	20	$Popul_3$	0	0	20

Table 5.1.1: Performance of average linkage hierarchical clustering in simulation

the one which optimizes that criterion is used. In our study, we have considered the average Mahalanobis distance between two different clusters and that within one cluster. The ratio of these two average distances was used as the criterion to optimize.

5.2.1 Model Assumption

In our study, we have considered model based clustering problems. To be more precise, we have assumed Gaussian clusters, i.e. the cluster means are i.i.d.

$$\boldsymbol{\mu}_i \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \mathbf{B}); \qquad i = 1, 2, \ldots, K \tag{5.2.1}$$

and the observations from the i^{th} cluster are i.i.d.

$$\mathbf{X}_{ij}|\boldsymbol{\mu}_i \sim \mathrm{N}_p(\boldsymbol{\mu}_i, \mathbf{W}); \qquad j = 1, 2, \ldots, N_i \tag{5.2.2}$$

where $\mathbf{B}$ and $\mathbf{W}$ are the between-group and within-group covariance matrices respectively.

In this chapter, we have assumed both $\mathbf{B}$ and $\mathbf{W}$ to be known. In Chapter 6, we will discuss the we will discuss the case where $\mathbf{B}$ and $\mathbf{W}$ are estimated from past data.

5.2.2 Criterion Function

Let d be the number of linear combinations to be obtained, and $\mathbf{L}$ be a $p \times d$ matrix such that $\mathbf{L}^T\mathbf{X}_{ij}$ is a vector of d linear combinations of the observation

vector $\mathbf{X}_{ij}$. We consider the class $\mathbb{L}$ of coefficient matrices each containing d many linear combinations, i.e.,

$$\mathbb{L} = \{\mathbf{L} : \mathbf{L} \text{ is } p \times d\}. \tag{5.2.3}$$

For any $\mathbf{L} \in \mathbb{L}$, $\mathbf{L}^T\mathbf{X}_{ij}|\boldsymbol{\mu}_i \sim \mathrm{N}(\mathbf{L}^T\boldsymbol{\mu}_i, \mathbf{L}^T\mathbf{W}\mathbf{L}) \forall i, j$. So, the expected Mahalanobis distance between two random observations from the same cluster, if $\mathbf{L}$ is used to transform the data, is

$$D_{ii}(\mathbf{L}) = \mathrm{E}\left[(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{ij'})^T(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{ij'})\right].$$

The expected Mahalanobis distance between two random observations from two different clusters, if $\mathbf{L}$ is used to transform the data, is

$$D_{ii'}(\mathbf{L}) = \mathrm{E}\left[(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{i'j'})^T(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{i'j'})\right].$$

We choose $\mathbf{L}_{\text{opt}} \in \mathbb{L}$ which maximizes the ratio between the two expected Mahalanobis distances , i.e.

$$\frac{D_{ii'}(\mathbf{L}_{\text{opt}})}{D_{ii}(\mathbf{L}_{\text{opt}})} = \max_{\mathbf{L}\in\mathbb{L}} \frac{D_{ii'}(\mathbf{L})}{D_{ii}(\mathbf{L})}, \tag{5.2.4}$$

where $i \neq i^{'}$ and $i, i^{'} \in \{1, 2, \ldots, K\}$. It is a reasonable and natural criterion to consider, and the set of 'optimum' linear combinations is expected to perform better than other such sets as discriminating directions.
The following proposition shows that maximizing (5.2.4) is equivalent to maximizing the criterion function

$$\Psi(\mathbf{L}) = \text{ trace}[(\mathbf{L}^T\mathbf{B}\mathbf{L})(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}] \tag{5.2.5}$$

with respect to $\mathbf{L} \in \mathbb{L}$.

Proposition 5.2.1. $\mathbf{L}_{opt} \in \mathbb{L}$ *satisfies*

$$\frac{D_{ii'}(\mathbf{L}_{opt})}{D_{ii}(\mathbf{L}_{opt})} = \max_{\mathbf{L}\in\mathbb{L}} \frac{D_{ii'}(\mathbf{L})}{D_{ii}(\mathbf{L})} \tag{5.2.6}$$

if and only if it satisfies

$$trace[(\mathbf{L}_{opt}^T\mathbf{B}\mathbf{L}_{opt})(\mathbf{L}_{opt}^T\mathbf{W}\mathbf{L}_{opt})^{-1}] = \max_{\mathbf{L}\in\mathbb{L}} \; trace[(\mathbf{L}^T\mathbf{B}\mathbf{L})(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}]. \tag{5.2.7}$$

Proof. We observe that

$$\begin{aligned}
D_{ii}(\mathbf{L}) =& \mathrm{E}\left[(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{ij'})^T(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{ij'})\right] \\
=& \mathrm{E}[(\mathbf{X}_{ij} - \mathbf{X}_{ij'})^T\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T(\mathbf{X}_{ij} - \mathbf{X}_{ij'})] \\
=& \mathrm{E}_{\mu_i}[\mathrm{E}\{(\mathbf{X}_{ij} - \mathbf{X}_{ij'})^T\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T(\mathbf{X}_{ij} - \mathbf{X}_{ij'})|\boldsymbol{\mu}_i\}] \\
=& \mathrm{E}_{\mu_i}[\ \mathrm{trace}\{\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T 2\mathbf{W}\}] \\
& \quad [\text{as } (\mathbf{X}_{ij} - \mathbf{X}_{ij'}) \sim \mathrm{N}(\mathbf{0}, 2\mathbf{W}) \\
& \quad \text{and } \mathbf{Y} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}) \Rightarrow \mathrm{E}(\mathbf{Y}^T\mathbf{A}\mathbf{Y}) = \mathrm{trace}(\mathbf{A}\boldsymbol{\Sigma})\ ] \\
=& 2\mathrm{E}_{\mu_i}[\ \mathrm{trace}\{(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{W}\mathbf{L}\}] \\
=& 2d, \qquad (5.2.8)
\end{aligned}$$

and

$$\begin{aligned}
D_{i,i'} =& \mathrm{E}\left[(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{i'j'})^T(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{X}_{ij} - \mathbf{L}^T\mathbf{X}_{i'j'})\right] \\
=& \mathrm{E}[(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})^T\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})] \\
=& \mathrm{E}_{\mu_i,\mu_{i'}}[\mathrm{E}\{(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})^T\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T(\mathbf{X}_{ij} - \mathbf{X}_{i'j'})|\boldsymbol{\mu}_i, \boldsymbol{\mu}_{i'}\}] \\
=& \mathrm{E}_{\mu_i,\mu_{i'}}[(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'})^T\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}) + \ \mathrm{trace}\{\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T 2\mathbf{W}\}] \\
& \qquad\qquad [\text{As } (\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) \sim \mathrm{N}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i'}, 2\mathbf{W})] \\
=& 2\ \mathrm{trace}\{\mathbf{L}(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{B}\} + 2\ \mathrm{trace}\{(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{W}\mathbf{L}\} \\
=& 2\ \mathrm{trace}\{(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\mathbf{L}^T\mathbf{B}\mathbf{L}\} + 2d. \qquad (5.2.9)
\end{aligned}$$

So, maximizing the ratio of the two average Mahalanobis distances (between-cluster and within-cluster) is equivalent to maximizing the trace of the ratio matrix $(\mathbf{L}^T\mathbf{B}\mathbf{L})(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}$. □

From another point of view also, we arrive at the same optimization criterion. If $\mathbf{L}$ is used to transform the data, the conditional distribution of the difference of two random observations $\mathbf{X}$ and $\mathbf{Y}$ from the clusters, is $\mathrm{N}(\mathbf{0}, \mathbf{L}^T\mathbf{W}\mathbf{L})$ or $\mathrm{N}(\mathbf{L}^T\boldsymbol{\mu}_i - \mathbf{L}^T\boldsymbol{\mu}_{i'}, \mathbf{L}^T\mathbf{W}\mathbf{L})$ depending on whether $\mathbf{X}, \mathbf{Y}$ come from the same cluster or different clusters, respectively. If we consider the average Mahalanobis distance between these two populations, that seems to be a reasonable quantity to maximize. The expected Mahalanobis distance is

$$\mathrm{E}[(\mathbf{L}^T\boldsymbol{\mu}_i - \mathbf{L}^T\boldsymbol{\mu}_{i'})^T(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}(\mathbf{L}^T\boldsymbol{\mu}_i - \mathbf{L}^T\boldsymbol{\mu}_{i'})] = \ \mathrm{trace}[(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}(\mathbf{L}^T\mathbf{B}\mathbf{L})], \qquad (5.2.10)$$

which is the quantity we obtained earlier.

5.2.3 Procedure for Choosing $\mathbf{L_{opt}}$

First, we observe that the criterion function is invariant under nonsingular transformations, i.e., $\Psi(\mathbf{L}) = \Psi(\mathbf{LC})$ for all $\mathbf{L} \in \mathbb{L}$ and for all nonsingular $\mathbf{C}$. In particular, if $\mathbf{L}_{\text{opt}} \in \mathbb{L}$ maximizes $\Psi(\cdot)$, $\mathbf{C}$ can be chosen suitably, namely

$$\mathbf{C} = \left(\mathbf{L}_{\text{opt}}^T \mathbf{W} \mathbf{L}_{\text{opt}}\right)^{-1/2},$$

such that $\mathbf{L}_0 = \mathbf{L}_{\text{opt}}\mathbf{C}$ satisfies $\mathbf{L}_0^T \mathbf{W} \mathbf{L}_0 = \mathbf{I}_d$ and $\mathbf{L}_0$ also maximizes $\Psi(\cdot)$. So, we can get a set of linear combinations $\mathbf{L}_0$ such that our criterion function is maximized, and at the same time, all the components of the transformed observations are independent, each having variance 1. So we concentrate on the class

$$\mathbb{L}_{\mathbf{W}} = \{\mathbf{L} \in \mathbb{L} : \mathbf{L}^T \mathbf{W} \mathbf{L} = \mathbf{I}_d\}. \tag{5.2.11}$$

For $\mathbf{L} \in \mathbf{L}_{\mathbf{W}}$, $\Psi(\mathbf{L}) = \text{trace}(\mathbf{L}^T \mathbf{B} \mathbf{L})$. In order to understand which choice of $\mathbf{L}_0 \in \mathbb{L}_{\mathbf{W}}$ would maximize $\text{trace}(\mathbf{L}^T \mathbf{B} \mathbf{L})$, it is better to consider the one-dimensional situation (i.e., $d = 1$) first and then extend the technique to higher dimensions. When $d = 1$, the above maximization problem is equivalent to maximizing $\frac{\mathbf{l}^T \mathbf{B} \mathbf{l}}{\mathbf{l}^T \mathbf{W} \mathbf{l}}$ with respect to $\mathbf{l} \in C(\mathbf{W})$. We know that if $\mathbf{W} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^T$ (where $\mathbf{P}$ is $p \times r$ and $\mathbf{\Lambda}$ is $r \times r$, $\text{rank}(\mathbf{W}) = r \leq p$) is the spectral decomposition of $\mathbf{W}$, and if we denote

$$\mathbf{W}^{-1/2} = \mathbf{P} \left(\Lambda\right)^{-1/2} \mathbf{P}^T \qquad (\text{where } \Lambda^{-1/2} \text{ is the sqare root of } \Lambda^{-1}), \tag{5.2.12}$$

then

$$\max_{\mathbf{l} \in C(\mathbf{W})} \frac{\mathbf{l}^T \mathbf{B} \mathbf{l}}{\mathbf{l}^T \mathbf{W} \mathbf{l}} = \lambda_{\max}(\mathbf{B}\mathbf{W}^+) = \lambda_{\max}(\mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2}) \tag{5.2.13}$$

and

$$\arg\max_{\mathbf{l} \neq \mathbf{0}} \frac{\mathbf{l}^T \mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2} \mathbf{l}}{\mathbf{l}^T \mathbf{l}} = \mathbf{l}_1 \Rightarrow \arg \max_{\mathbf{l} \in C(\mathbf{W})} \frac{\mathbf{l}^T \mathbf{B} \mathbf{l}}{\mathbf{l}^T \mathbf{W} \mathbf{l}} = \mathbf{W}^{-1/2} \mathbf{l}_1, \tag{5.2.14}$$

as $\mathbf{l}_1 \in C(\mathbf{W}) = C(\mathbf{P})$. So, it is natural to expect that, if we consider the spectral decomposition $\mathbf{Q}\mathbf{D}\mathbf{Q}^T$ of $\mathbf{W}^{-1/2} \mathbf{B} \mathbf{W}^{-1/2}$, choose the first d columns of $\mathbf{Q}$ and multiply each of them by $\mathbf{W}^{-1/2}$, we should get our desired $\mathbf{L}_0$. The following propositions show that indeed we get the optimizer of $\text{trace}(\mathbf{L}^T \mathbf{B} \mathbf{L})$ in the class $\mathbb{L}_{\mathbf{W}}$ by the above procedure.

Proposition 5.2.2. *If* $\mathbf{D}$ *is a diagonal matrix with diagonal entries* $d_1 \geq d_2 \geq \ldots \geq d_n$*, then*

$$\max_{\{\mathbf{L}_{n\times r}:\mathbf{L}^T\mathbf{L}=\mathbf{I}_r\}} \textit{trace}\left(\mathbf{L}^T\mathbf{D}\mathbf{L}\right) = \sum_{i=1}^{r} d_i \qquad \forall r, 1 \leq r \leq n. \tag{5.2.15}$$

Equality holds when $\mathbf{L}$ *consists of first* r *canonical eigenvectors.*

Proof. Let $\mathbf{L} = [\mathbf{l}_1, \ldots, \mathbf{l}_n]^T$.

If $\mathbf{L}^T\mathbf{L} = \mathbf{I}_r$, largest eigenvalue of $\mathbf{L}\mathbf{L}^T$ is 1.

So the i^{th} diagonal entry of $\mathbf{L}\mathbf{L}^T$, $\mathbf{l}_i^T\mathbf{l}_i \leq 1 \forall i = 1, 2, \ldots n$ and $\sum_{i=1}^{n} \mathbf{l}_i^T\mathbf{l}_i = \text{trace}\ (\mathbf{L}\mathbf{L}^T) = r$. Now

$$\begin{aligned}
\text{trace}\left(\mathbf{L}^T\mathbf{D}\mathbf{L}\right) &= \text{trace}\left(\mathbf{D}\mathbf{L}\mathbf{L}^T\right) \\
&= \sum_{i=1}^{n} d_i \mathbf{l}_i^T\mathbf{l}_i \\
&\leq \sum_{i=1}^{r} d_i \mathbf{l}_i^T\mathbf{l}_i + d_r \sum_{i=r+1}^{n} \mathbf{l}_i^T\mathbf{l}_i \\
&= \sum_{i=1}^{r} d_i \mathbf{l}_i^T\mathbf{l}_i + d_r(r - \sum_{i=1}^{r} \mathbf{l}_i^T\mathbf{l}_i) \\
&= \sum_{i=1}^{r} (d_i - d_r)\mathbf{l}_i^T\mathbf{l}_i + r.d_r \\
&= \sum_{i=1}^{r} (d_i - d_r) + r.d_r \\
&= \sum_{i=1}^{r} d_i. \qquad (5.2.16)
\end{aligned}$$

Equality holds when $\mathbf{L}$ consists of first r unit vectors. □

Corollary 5.2.3. *For any n.n.d matrix* $\mathbf{M}$*, under the restriction* $\mathbf{L}^T\mathbf{L} = \mathbf{I}_r$*, trace of* $\mathbf{L}^T\mathbf{M}\mathbf{L}$ *is maximized when the eigenvectors of* $\mathbf{M}$ *corresponding to its largest* r *many eigenvalues are used.*

Proof. Consider the spectral decomposition of $\mathbf{M}$, say $\mathbf{M} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$. For any $\mathbf{L}$, with $\mathbf{L}^T\mathbf{L} = \mathbf{I}_r$, $\mathbf{L}^T\mathbf{Q}\mathbf{Q}^T\mathbf{L} = \mathbf{L}^T\mathbf{L} = \mathbf{I}_r$. So,

$$\text{trace}(\mathbf{L}^T\mathbf{M}\mathbf{L}) = \text{trace}(\mathbf{L}^T\mathbf{Q}\mathbf{D}\mathbf{Q}^T\mathbf{L}) \leq \sum_{i=1}^{r} d_i \quad \text{(by Proposition 5.2.2).} \tag{5.2.17}$$

Equality holds if $\mathbf{Q}^T\mathbf{L} = [\mathbf{I}_r\mathbf{0}]^T$, or equivalently $\mathbf{L} = \mathbf{Q}[\mathbf{I}_r\mathbf{0}]^T$, i.e., $\mathbf{L}$ consists of eigenvectors of $\mathbf{M}$ corresponding to its largest r eigenvalues. □

Proposition 5.2.4. *For any two n.n.d matrices* $\mathbf{B}$ *and* $\mathbf{W}$*, under the restriction* $\mathbf{L}^T\mathbf{W}\mathbf{L} = \mathbf{I}_d$*, trace of* $\mathbf{L}^T\mathbf{B}\mathbf{L}$ *is maximized at* $\mathbf{L}_0 = \mathbf{W}^{-1/2}\mathbf{L}_1$*, where* $\mathbf{L}_1$ *consists of orthogonal eigenvectors of* $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ *corresponding to its* d *many largest eigenvalues.*

Proof. Let the eigenvalues of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ be $\lambda_1 \geq \ldots \geq \lambda_p$. We note that

$$\mathbf{L}^T\mathbf{W}\mathbf{L} = \mathbf{I}_d \Rightarrow \mathbf{L}^T\mathbf{P}\mathbf{\Lambda}\mathbf{P}^T\mathbf{L} = \mathbf{I}_d \tag{5.2.18}$$

$$\text{(where } \mathbf{W} = \mathbf{P}_{p\times r}\mathbf{\Lambda}_{r\times r}\mathbf{P}^T \text{ is the spectral decomposition.)}$$

$$\Rightarrow \mathbf{u}^T\mathbf{u} = \mathbf{I}_d \qquad \text{(where } \mathbf{L} = \mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{u}\text{, assuming } \mathbf{L} \in C(\mathbf{W})\text{)} \tag{5.2.19}$$

and

$$\mathbf{L}^T\mathbf{B}\mathbf{L} = \mathbf{u}^T\mathbf{\Lambda}^{-1/2}\mathbf{P}^T\mathbf{B}\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{u}. \tag{5.2.20}$$

So,

$$\begin{aligned}
&\max_{\mathbf{L}\in\mathbb{L}_\mathbf{W}} \text{trace}(\mathbf{L}^T\mathbf{B}\mathbf{L}) = \max_{\mathbf{u}^T\mathbf{u}=\mathbf{I}_d} \text{trace}(\mathbf{u}^T\mathbf{\Lambda}^{-1/2}\mathbf{P}^T\mathbf{B}\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{u}) \\
&= \text{ sum of the largest } d \text{ eigenvalues of } \mathbf{\Lambda}^{-\frac{1}{1}}\mathbf{P}^T\mathbf{B}\mathbf{P}\mathbf{\Lambda}^{-1/2} \\
&= \text{ sum of the largest } d \text{ eigenvalues of } \mathbf{P}\mathbf{\Lambda}^{-\frac{1}{1}}\mathbf{P}^T\mathbf{B}\mathbf{P}\mathbf{\Lambda}^{-1/2}\mathbf{P}^T \qquad (\text{as } \mathbf{P}^T\mathbf{P} = \mathbf{I}_r) \\
&= \text{ sum of the largest } d \text{ eigenvalues of } \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2} = \sum_{i=1}^{d}\lambda_i. \qquad (5.2.21)
\end{aligned}$$

If $\mathbf{L}_1 = [\mathbf{l}_1, \ldots, \mathbf{l}_d]$ consists of the orthogonal eigenvectors of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ corresponding to $\lambda_1, \ldots, \lambda_d$, then $\mathbf{L}_1^T\mathbf{L}_1 = \mathbf{I}_d$ and $\mathbf{l}_i^T\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{l}_i = \lambda_i$ for $i = 1, \ldots, d$. Now, if we take $\mathbf{L}_0 = \mathbf{W}^{-1/2}\mathbf{L}_1$, then

$$\begin{aligned}
\mathbf{L}_0^T\mathbf{W}\mathbf{L}_0 &= \mathbf{L}_1^T\mathbf{W}^{-1/2}\mathbf{W}\mathbf{W}^{-1/2}\mathbf{L}_1 \\
&= \mathbf{L}_1^T\mathbf{P}\mathbf{P}^T\mathbf{L}_1 \\
&= \mathbf{L}_1^T\mathbf{L}_1 \qquad (\text{as each } \mathbf{l}_i \in C(\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}) \subseteq C(\mathbf{W}^{-1/2}) = C(\mathbf{P})) \\
&= \mathbf{I}_d, \qquad (5.2.22)
\end{aligned}$$

i.e., $\mathbf{L}_0 \in \mathbb{L}_\mathbf{W}$ and

$$\text{trace}(\mathbf{L}_0^T\mathbf{B}\mathbf{L}_0) = \sum_{i=1}^{d}\mathbf{l}_i^T\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{l}_i = \sum_{i=1}^{d}\lambda_i. \tag{5.2.23}$$

Hence $\mathbf{L}_0$ maximizes trace of $\mathbf{L}^T\mathbf{B}\mathbf{L}$. □

So, from Proposition 5.2.4, we conclude that if $\mathbf{B}$ and $\mathbf{W}$ are known, in order to maximize the trace of $\left[(\mathbf{L}^T\mathbf{B}\mathbf{L})(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\right]$ with respect to $\mathbf{L} \in \mathbb{L}$, it is enough to consider $\mathbf{L}_0 = \mathbf{W}^{-1/2}\mathbf{L}_1$, where $\mathbf{L}_1$ consists of the orthogonal eigenvectors of the ratio matrix $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ corresponding to its leading d many eigenvalues. As $\mathbf{L}_0 \in \mathbb{L}_{\mathbf{W}}$, it is also ensured that all the components of the transformed observations are conditionally independent, each having conditional variance 1. So, we may use the Euclidian distance. If we had considered some other optimal linear combinations, we would have had to consider the Mahalanobis distance, which is computationally more costly. Moreover,

$$\mathbf{L}_0^T\mathbf{B}\mathbf{L}_0 = \mathbf{L}_1^T\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{L}_1 = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_d \end{bmatrix},$$

where $\lambda_1, \lambda_2, \dots, \lambda_d$ are leading eigenvalues of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ with $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$. This means that if $\mathbf{L}_0$ is used to transform the data, the transformed cluster means become independent with decreasing variances.

In general, if $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ is the spectral decomposition of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, where

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_r \end{bmatrix},$$

the assertions given in Table 5.2.1 can be concluded.

Once $\mathbf{L}$ is chosen as $\mathbf{L}_0$, the function $\Psi(\cdot)$ defined in (5.2.5) reduces to $\lambda_1 + \lambda_2 + \cdots + \lambda_d$. Since this is a monotonically increasing function of d, it cannot be used for selecting the number of linear combinations. We have to look for some other criterion.

5.3 Number of Linear Combinations to Choose

In the last section, we have seen that if $\mathbf{B}$ and $\mathbf{W}$ are known, then one can optimally select d linear combinations of the variables for model-based clustering. However, the number of linear combinations (d) has to be specified. In this section, we have addressed the important and difficult problem of choosing the optimal number of linear combinations.

The data can be transformed suitably such that

• ***the transformed within-group variation matrix is the identity matrix i.e., all components of the transformed observations are conditionally independent with conditional variance 1;***

• ***the transformed between-group variation matrix is a diagonal matrix with decreasing diagonal entries i.e., all components of the transformed cluster means are independent with decreasing variances; and***

• ***for any*** d, $1 \leq d \leq r$, ***the first*** d ***components of the transformed observations correspond to the optimal*** d ***many linear comb--inations according to the optimality criterion (5.2.5).***

Table 5.2.1: Summary of section 5.2

5.3.1 Criterion for Choosing

We need to develop a suitable criterion in order to compare the performance of the optimal sets of linear combinations corresponding to several d's. In view of Table 5.2.1, the within-group dispersion matrix $\mathbf{W}$ can be assumed to be the identity matrix $\mathbf{I}_p$, and the between-group dispersion matrix $\mathbf{B}$ can be assumed to be a diagonal matrix with decreasing diagonal entries, say

$$\mathbf{B} = \begin{bmatrix} b_1 & 0 & \dots & 0 \\ 0 & b_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & b_p \end{bmatrix}, \quad \text{where} \quad b_1 \geq b_2 \geq \dots \geq b_p. \tag{5.3.1}$$

We have to choose one among the p many competing subsets of variables, namely the subsets consisting of the first d many variables ($1 \leq d \leq p$). Since $\mathbf{W} = \mathbf{I}_p$, Euclidean distance between the observations is a reasonable measure of dissimilarity.

Consider a hypothetical data set where group identities of all the clusters are known. If the observations are ordered such that the observations corresponding to the i^{th} cluster are placed ahead of those corresponding to the j^{th} cluster whenever $i < j$, then the distance matrix of the observations has the structure shown in Figure 5.3.1. The expected distance matrix shown in Figure 5.3.2 has the interesting feature that the between-cluster entries always exceed the within-cluster ones. Any reasonable clustering method used

on the (unconditional) expected distance matrix would yield correct clusters, irrespective of which set of variables is used.

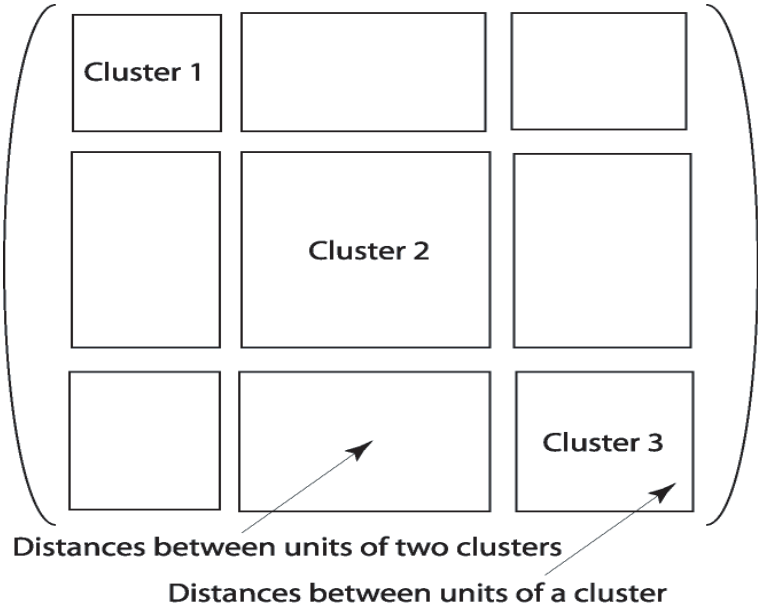


Figure 5.3.1: Partitioned Distance Matrix1

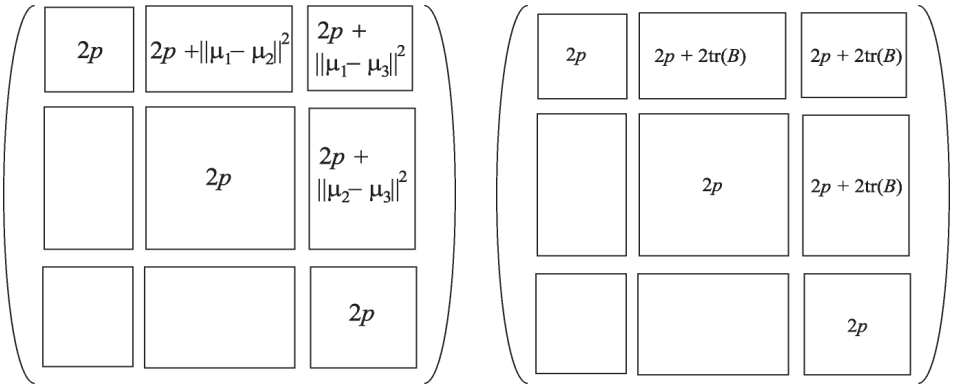
In practice, a random distance matrix does not have the above property. However, comparison with the expected distance matrix suggests that a distance matrix will be more amenable for correct clustering, if between-cluster distances are generally greater than the within-cluster distances with high probability.
Keeping this in mind, we have considered the probability of the random event that between-cluster Euclidean distance is larger than within-cluster Euclidean distance, as the criterion function for selecting d. To be more precise, if $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, \ldots, X_{ijp})^T$ denote the j^{th} observation from the i^{th} cluster and if

$$\mathbf{X}_{ij}^d = (X_{ij1}, X_{ij2}, \ldots, X_{ijd})^T, \tag{5.3.2}$$

then, we take our criterion function $\Gamma(d)$ as

$$\Gamma(d) = \mathrm{P}\left[\left\|\mathbf{X}_{11}^d - \mathbf{X}_{21}^d\right\|^2 > \left\|\mathbf{X}_{11}^d - \mathbf{X}_{12}^d\right\|^2\right]. \tag{5.3.3}$$

In the above expression, $\mathbf{X}_{11}^d$ occurs in both sides of the inequality. This corresponds to elements occurring on the same row or column of the matrix



(a) (b)

Figure 5.3.2: (a) Expected distance matrix – conditional on the cluster means (b) Expected distance matrix – unconditional

of Figure 5.3.1. For comparison of other pairs of elements of the distance matrix, we consider the additional criterion function

$$\Gamma'(d) = \mathrm{P}\left[\left\|\mathbf{X}_{11}^{d} - \mathbf{X}_{21}^{d}\right\|^2 > \left\|\mathbf{X}_{13}^{d} - \mathbf{X}_{12}^{d}\right\|^2\right]. \tag{5.3.4}$$

We may maximize $\Gamma(d)$ and $\Gamma'(d)$ with respect to d, and choose the smaller of the two maximizers for parsimony.
Note that, the *expected* distance between elements of different clusters is *always* more than that between elements of the same cluster, and the (positive) difference is a non-decreasing function of the number of dimensions. Therefore, the *expected* distance matrix could not possibly be used for selecting d. In the next section, we demonstrate that the suggested criterion has an in-built penalty for redundant dimensions.

5.3.2 Ill Effects of Nondiscriminating Variables

We now proceed to give a theoretical confirmation of the notion that variables which do not differ across clusters, can only make the distance matrix less suitable for correct clustering – in a probabilistic sense. We show that $\Gamma(\cdot)$ and $\Gamma'(\cdot)$ decrease if such noisy variables are included in the clustering procedure. The main result – Proposition 5.3.5 – would follow from the following proposition.

Proposition 5.3.1. *If for* $i = 1, 2$ *and* $j = 1, 2, 3$

$$\begin{pmatrix}\mathbf{X}_{ij}\\ \mathbf{Y}_{ij}\end{pmatrix} \sim N\left[\begin{pmatrix}\boldsymbol{\mu}_i\\ \boldsymbol{\nu}\end{pmatrix}, \begin{pmatrix}\mathbf{I}_d & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_r\end{pmatrix}\right]$$

then

$$\begin{aligned}
(1)\quad & P\left[\|\mathbf{X}_{21}-\mathbf{X}_{11}\|^2 > \|\mathbf{X}_{12}-\mathbf{X}_{13}\|^2\right] \\
> \quad & P\left[\left\|\begin{pmatrix}\mathbf{X}_{21}\\ \mathbf{Y}_{21}\end{pmatrix}-\begin{pmatrix}\mathbf{X}_{11}\\ \mathbf{Y}_{11}\end{pmatrix}\right\|^2 > \left\|\begin{pmatrix}\mathbf{X}_{12}\\ \mathbf{Y}_{12}\end{pmatrix}-\begin{pmatrix}\mathbf{X}_{13}\\ \mathbf{Y}_{13}\end{pmatrix}\right\|^2\right]; \\
(2)\quad & P\left[\|\mathbf{X}_{21}-\mathbf{X}_{11}\|^2 > \|\mathbf{X}_{12}-\mathbf{X}_{11}\|^2\right] \\
> \quad & P\left[\left\|\begin{pmatrix}\mathbf{X}_{21}\\ \mathbf{Y}_{21}\end{pmatrix}-\begin{pmatrix}\mathbf{X}_{11}\\ \mathbf{Y}_{11}\end{pmatrix}\right\|^2 > \left\|\begin{pmatrix}\mathbf{X}_{12}\\ \mathbf{Y}_{12}\end{pmatrix}-\begin{pmatrix}\mathbf{X}_{11}\\ \mathbf{Y}_{11}\end{pmatrix}\right\|^2\right].
\end{aligned}$$

In order to prove the stated result, we need to develop a series of lemmas that follow.

Let $\mathfrak{X} = \{X : X \text{ has density } f(\cdot) \text{ satisfying } f(x) > f(-x)\ \forall\ x > 0\}$.

Lemma 5.3.2. *If $X, Y \in \mathfrak{X}$ and they are independent, then $XY \in \mathfrak{X}$.*

Proof. Suppose that X and Y have densities $f(\cdot)$ and $g(\cdot)$, respectively. If XY has density $h(\cdot)$, then

$$h(u) = \int_{-\infty}^{\infty} f\left(\frac{u}{s}\right)\cdot g(s)\frac{1}{|s|}\,ds = \int_0^{\infty}\left[f\left(\frac{u}{s}\right)\cdot g(s) + f\left(-\frac{u}{s}\right)\cdot g(-s)\right]\frac{1}{|s|}\,ds.$$

So, for any $u > 0$,

$$h(u) - h(-u) = \int_0^{\infty}\left[f\left(\frac{u}{s}\right) - f\left(-\frac{u}{s}\right)\right]\cdot[g(s) - g(-s)]\,\frac{1}{|s|}\,ds > 0$$

as $X, Y \in \mathfrak{X}$. Hence, $XY \in \mathfrak{X}$. □

Lemma 5.3.3. *If*

(1) $X \in \mathfrak{X}$,
(2) Y is unimodal and symmetric about 0 with support $\mathbb{R}$, and
(3) X and Y are independent,

then, $X + Y \in \mathfrak{X}$.

Proof. Suppose that X and Y have densities $f(\cdot)$ and $g(\cdot)$, respectively. Then, $f(x) > f(-x)\ \forall\ x > 0$ and $g(y_1) \geq g(y_2)$, whenever $0 \leq y_1 < y_2$. If $X + Y$ has density $h(\cdot)$, then

$$h(u) = \int_{-\infty}^{\infty} f(s)\cdot g(u-s)ds = \int_0^{\infty}\left[f(s)\cdot g(u-s) + f(-s)\cdot g(u+s)\right]ds,$$

and

$$h(-u) = \int_0^\infty [f(s) \cdot g(-u-s) + f(-s) \cdot g(-u+s)]\, ds$$
$$= \int_0^\infty [f(s) \cdot g(u+s) + f(-s) \cdot g(u-s)]\, ds.$$

So, for any $u > 0$,

$$h(u) - h(-u) = \int_0^\infty [f(s) - f(-s)] \cdot [g(u-s) - g(u+s)] ds > 0,$$

as $g(u-s) = g(|u-s|) \geq g(u+s)$ and $f(s) > f(-s)$ for all $u, s > 0$. □

Lemma 5.3.4. *If*

(1) $X \in \mathfrak{X}$,
(2) Y *is symmetric about 0, and*
(3) X *and* Y *are independent,*

then, $P[X > 0] > P[X + Y > 0]$.

Proof. Suppose that X has density $f(\cdot)$ with distribution function $F(\cdot)$. Then,

$$[F(x) + F(-x)]' = f(x) - f(-x) \begin{cases} > 0 & \text{if } x > 0 \\ = 0 & \text{if } x = 0 \\ < 0 & \text{if } x < 0. \end{cases}$$

So, 0 is the global minimum of the function $F(x) + F(-x)$. Now, if $g(\cdot)$ is

the density of Y, then

$$\begin{aligned}
P[X+Y>0] &= \int_{-\infty}^{\infty} P[X+y>0|Y=y]\cdot g(y)dy \\
&= \int_{-\infty}^{\infty} P[X+y>0]\cdot g(y)dy \qquad \text{(using independence)} \\
&= 1-\int_{-\infty}^{\infty} P[X+y\leq 0]\cdot g(y)dy \\
&= 1-\int_{-\infty}^{\infty} F(-y)\cdot g(y)dy \\
&= 1-\int_{0}^{\infty} [F(y)+F(-y)]\cdot g(y)dy \qquad [\text{as } g(-y)=g(y)\forall y] \\
&< 1-\int_{0}^{\infty} 2F(0)\cdot g(y)dy \qquad \text{(using the global minimum)} \\
&= 1-2F(0).\frac{1}{2} \qquad \text{(using symmetry of } g(\cdot) \text{ about 0)} \\
&= 1-F(0) = P[X>0].
\end{aligned}$$

Hence, the inequality follows. □

Using the above Lemmas we can prove Proposition 5.3.1.

Proof of Proposition 5.3.1. Suppose that $\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. Then

$$\begin{pmatrix} \mathbf{W}^1 \\ \mathbf{Z}^1 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{21} - \mathbf{X}_{11} \\ \mathbf{Y}_{21} - \mathbf{Y}_{11} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} 2\mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & 2\mathbf{I}_r \end{pmatrix} \right]$$

and

$$\begin{pmatrix} \mathbf{W}^2 \\ \mathbf{Z}^2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_{12} - \mathbf{X}_{13} \\ \mathbf{Y}_{12} - \mathbf{Y}_{13} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mathbf{0}_d \\ \mathbf{0}_r \end{pmatrix}, \begin{pmatrix} 2\mathbf{I}_d & \mathbf{0} \\ \mathbf{0} & 2\mathbf{I}_r \end{pmatrix} \right],$$

with

$$\mathrm{cov}\left[\begin{pmatrix} \mathbf{W}^1 \\ \mathbf{Z}^1 \end{pmatrix}, \begin{pmatrix} \mathbf{W}^2 \\ \mathbf{Z}^2 \end{pmatrix} \right] = \mathbf{0}_{d+r,d+r}.$$

Let $\mathbf{A}$ be a $d \times d$ orthogonal matrix, such that the first row is proportional to $\boldsymbol{\mu}$. Now, if

$$\begin{pmatrix} \mathbf{U}^k \\ \mathbf{V}^k \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_r \end{pmatrix} \begin{pmatrix} \mathbf{W}^k \\ \mathbf{Z}^k \end{pmatrix}, \qquad \text{for } k = 1, 2$$

then

$$\begin{pmatrix}\mathbf{U}^1\\ \mathbf{V}^1\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\begin{pmatrix}\delta\\ \mathbf{0}_{d-1}\end{pmatrix}\\ \mathbf{0}_r\end{pmatrix}, \begin{pmatrix}2\mathbf{I}_d & \mathbf{0}\\ \mathbf{0} & 2\mathbf{I}_r\end{pmatrix}\right]; \quad \begin{pmatrix}\mathbf{U}^2\\ \mathbf{V}^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}2\mathbf{I}_d & \mathbf{0}\\ \mathbf{0} & 2\mathbf{I}_r\end{pmatrix}\right]$$

with $\delta = \|\boldsymbol{\mu}\|$ and

$$\mathrm{cov}\left[\begin{pmatrix}\mathbf{U}^1\\ \mathbf{V}^1\end{pmatrix}, \begin{pmatrix}\mathbf{U}^2\\ \mathbf{V}^2\end{pmatrix}\right] = \mathbf{0}_{d+r,d+r}.$$

Since $\left\|\mathbf{U}^k\right\|^2 = \left\|W^k\right\|^2$ and $\mathbf{V}^k = \mathbf{Z}^k$ for $k = 1, 2$, we need to show that

$$\mathrm{P}\left[\left\|\mathbf{U}^1\right\|^2 - \left\|\mathbf{U}^2\right\|^2 > 0\right] > \mathrm{P}\left[\left\|\mathbf{U}^1\right\|^2 - \left\|\mathbf{U}^2\right\|^2 + \left\|\mathbf{V}^1\right\|^2 - \left\|\mathbf{V}^2\right\|^2 > 0\right] \tag{5.3.5}$$

Now, we observe that

$$\begin{pmatrix}\mathbf{U}_1^1\\ \mathbf{U}_1^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\delta\\ 0\end{pmatrix}, \begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}\right].$$

Hence,

$$\begin{pmatrix}\mathbf{U}_1^1 + \mathbf{U}_1^2\\ \mathbf{U}_1^1 - \mathbf{U}_1^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\delta\\ \delta\end{pmatrix}, \begin{pmatrix}4 & 0\\ 0 & 4\end{pmatrix}\right].$$

It is easy to see that $\mathbf{U}_1^1 + \mathbf{U}_1^2, \mathbf{U}_1^1 - \mathbf{U}_1^2 \in \mathfrak{X}$ as $\delta > 0$. So, by Lemma 5.3.2, we get

$$(\mathbf{U}_1^1)^2 - (\mathbf{U}_1^2)^2 = (\mathbf{U}_1^1 + \mathbf{U}_1^2)(\mathbf{U}_1^1 - \mathbf{U}_1^2) \in \mathfrak{X}.$$

Next, we observe that for $k = 2, \ldots, d$,

$$\begin{pmatrix}\mathbf{U}_k^1\\ \mathbf{U}_k^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}\right] \quad \text{and} \quad \begin{pmatrix}\mathbf{U}_k^1 + \mathbf{U}_k^2\\ \mathbf{U}_k^1 - \mathbf{U}_k^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}4 & 0\\ 0 & 4\end{pmatrix}\right],$$

which imply that for each $k = 2, \ldots, d$, $(\mathbf{U}_k^1)^2 - (\mathbf{U}_k^2)^2$ is symmetric about 0 and is unimodal and hence so is $\sum_{k=2}^d [(\mathbf{U}_k^1)^2 - (\mathbf{U}_k^2)^2]$. Also its support is $\mathbb{R}$. Hence, by using Lemma 5.3.3, it can be concluded that $\|\mathbf{U}^1\|^2 - \|\mathbf{U}^2\|^2 = \sum_{k=1}^d [(\mathbf{U}_k^1)^2 - (\mathbf{U}_k^2)^2] \in \mathfrak{X}$. Finally, since $(\mathbf{V}^1, \mathbf{V}^2)$ has the same distribution as $(\mathbf{V}^2, \mathbf{V}^1)$, it follows that $\|\mathbf{V}^1\|^2 - \|\mathbf{V}^2\|^2$ is symmetric about 0, and is independent of $\|\mathbf{U}^1\|^2 - \|\mathbf{U}^2\|^2$. Hence, (5.3.5) follows from Lemma 5.3.4.

The second assertion follows by a similar argument. If we take $\boldsymbol{\mu}$ and $\mathbf{A}$ as in the earlier case, and define

$$\begin{pmatrix}\mathbf{U}^1\\ \mathbf{V}^1\end{pmatrix} = \begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_r\end{pmatrix}\begin{pmatrix}\mathbf{X}_{21} - \mathbf{X}_{11}\\ \mathbf{Y}_{21} - \mathbf{Y}_{11}\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}\mathbf{U}^2\\ \mathbf{V}^2\end{pmatrix} = \begin{pmatrix}\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_r\end{pmatrix}\begin{pmatrix}\mathbf{X}_{12} - \mathbf{X}_{11}\\ \mathbf{Y}_{12} - \mathbf{Y}_{11}\end{pmatrix},$$

then

$$\begin{pmatrix}\mathbf{U}^1\\ \mathbf{V}^1\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\begin{pmatrix}\delta\\ \mathbf{0}_{d-1}\end{pmatrix}\\ \mathbf{0}_r\end{pmatrix}, \begin{pmatrix}2\mathbf{I}_d & \mathbf{0}\\ \mathbf{0} & 2\mathbf{I}_r\end{pmatrix}\right] \quad \begin{pmatrix}\mathbf{U}^2\\ \mathbf{V}^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}2\mathbf{I}_d & \mathbf{0}\\ \mathbf{0} & 2\mathbf{I}_r\end{pmatrix}\right]$$

with $\delta = \|\boldsymbol{\mu}\|$ and

$$\mathrm{cov}\left[\begin{pmatrix}\mathbf{U}^1\\ \mathbf{V}^1\end{pmatrix}, \begin{pmatrix}\mathbf{U}^2\\ \mathbf{V}^2\end{pmatrix}\right] = \mathbf{I}_{d+r}.$$

As earlier, we need to show that equation (5.3.5) holds. In this case,

$$\begin{pmatrix}\mathbf{U}_1^1\\ \mathbf{U}_1^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\delta\\ 0\end{pmatrix}, \begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\right].$$

Hence,

$$\begin{pmatrix}\mathbf{U}_1^1 + \mathbf{U}_1^2\\ \mathbf{U}_1^1 - \mathbf{U}_1^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\delta\\ \delta\end{pmatrix}, \begin{pmatrix}6 & 0\\ 0 & 2\end{pmatrix}\right].$$

It is easy to see that $\mathbf{U}_1^1 + \mathbf{U}_1^2, \mathbf{U}_1^1 - \mathbf{U}_1^2 \in \mathfrak{X}$ as $\delta > 0$. So, by Lemma 5.3.2,

$$(\mathbf{U}_1^1)^2 - (\mathbf{U}_1^2)^2 = (\mathbf{U}_1^1 + \mathbf{U}_1^2)(\mathbf{U}_1^1 - \mathbf{U}_1^2) \in \mathfrak{X}.$$

Next, we observe that for $k = 2, \ldots, d$,

$$\begin{pmatrix}\mathbf{U}_k^1\\ \mathbf{U}_k^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\right] \quad \text{and} \quad \begin{pmatrix}\mathbf{U}_k^1 + \mathbf{U}_k^2\\ \mathbf{U}_k^1 - \mathbf{U}_k^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}6 & 0\\ 0 & 2\end{pmatrix}\right],$$

which imply that for each $k = 2, \ldots, d$, $(\mathbf{U}_k^1)^2 - (\mathbf{U}_k^2)^2$ is symmetric about 0 and is unimodal. Since symmetry and unimodality are preserved under convolution, $\sum_{k=2}^{d}[(\mathbf{U}_k^1)^2 - (\mathbf{U}_k^2)^2]$ is symmetric about 0 and is unimodal with support $\mathbb{R}$. Again, for each $k = 1, \ldots, d$,

$$\begin{pmatrix}\mathbf{V}_k^1\\ \mathbf{V}_k^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}2 & 1\\ 1 & 2\end{pmatrix}\right] \quad \text{and} \quad \begin{pmatrix}\mathbf{V}_k^1 + \mathbf{V}_k^2\\ \mathbf{U}_k^1 - \mathbf{U}_k^2\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}6 & 0\\ 0 & 2\end{pmatrix}\right],$$

which imply that $(\mathbf{V}_k^1)^2 - (\mathbf{V}_k^2)^2$ is symmetric about 0 and is unimodal. Hence, by an argument similar to the one used earlier, $\|\mathbf{V}^1\|^2 - \|\mathbf{V}^2\|^2 = \sum_{k=1}^{d}[(\mathbf{V}_k^1)^2 - (\mathbf{V}_k^2)^2]$ is symmetrically distributed about 0. Finally, applying Lemma 5.3.3 to $(\mathbf{U}_k^1)^2 - (\mathbf{U}_k^2)^2$ and $\sum_{k=2}^{d}[(\mathbf{U}_k^1)^2 - (\mathbf{U}_k^2)^2]$, we see that $\|\mathbf{U}^1\|^2 - \|\mathbf{U}^2\|^2 \in \mathfrak{X}$. The result (5.3.5) follows by applying Lemma 5.3.4 to $\|\mathbf{U}^1\|^2 - \|\mathbf{U}^2\|^2$ and $\|\mathbf{V}^1\|^2 - \|\mathbf{V}^2\|^2$. □

Proposition 5.3.1 leads to the main result of this section.

Proposition 5.3.5. *Suppose that the $\boldsymbol{\mu}_i$'s are as in (5.2.1) and the $\mathbf{X}_{ij}$'s are as in (5.2.2) with $\mathbf{W} = \mathbf{I}_p$ and $\mathbf{B}$ as in (5.3.1) such that $b_k = 0$ for some $k, 1 < k \leq p$. Then,*

$$\Gamma(k-1) > \Gamma(k) > \Gamma(k+1) > \ldots > \Gamma(p), \tag{5.3.6}$$

and

$$\Gamma^{'}(k-1) > \Gamma^{'}(k) > \Gamma^{'}(k+1) > \ldots > \Gamma^{'}(p). \tag{5.3.7}$$

Moreover, if $b_k = 0$ for some k, then $\Gamma(p) \downarrow 1/2$ and $\Gamma^{'}(p) \downarrow 1/2$ as $p \to \infty$.

Proof. If $b_k = 0$ for some $k, 1 < k \leq p$, then

$$\boldsymbol{\mu}_i = \begin{pmatrix} \boldsymbol{\eta}_i \\ \boldsymbol{\nu} \end{pmatrix} \forall i, \quad \text{where } \boldsymbol{\nu} \text{ is constant vector of length } p-k+1.$$

Now, if $\mathbf{X}_{ij} = \left(\mathbf{X}_{ij}^{k-1}, \mathbf{Y}_{ij}\right)^T$, then

$$\begin{pmatrix} \mathbf{X}_{ij}^{k-1} \\ \mathbf{Y}_{ij} \end{pmatrix} \Bigg| \begin{pmatrix} \boldsymbol{\eta}_i \\ \boldsymbol{\nu} \end{pmatrix} \sim \mathrm{N}_p \left[\begin{pmatrix} \boldsymbol{\eta}_i \\ \boldsymbol{\nu} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p-k+1} \end{pmatrix} \right], \tag{5.3.8}$$

where $\mathbf{X}_{ij}^{k-1}$ and $\mathbf{Y}_{ij}^{l}$ are as in 5.3.2. Now, for any $l, 1 \leq l \leq p-K+1$, applying Proposition 5.3.1(b) to $\left(\mathbf{X}_{ij}^{k-1}, \mathbf{Y}_{ij}^{l}\right)^T | \boldsymbol{\mu}_i$ for $i, j = 1, 2$, we get

$$\begin{aligned} &\mathrm{P}\left[\left\| \begin{pmatrix} \mathbf{X}_{21}^{k-1} \\ \mathbf{Y}_{21}^{l-1} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{11}^{k-1} \\ \mathbf{Y}_{11}^{l-1} \end{pmatrix} \right\|^2 > \left\| \begin{pmatrix} \mathbf{X}_{12}^{k-1} \\ \mathbf{Y}_{12}^{l-1} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{11}^{k-1} \\ \mathbf{Y}_{11}^{l-1} \end{pmatrix} \right\|^2 \Bigg| \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \right] \\ >&\mathrm{P}\left[\left\| \begin{pmatrix} \mathbf{X}_{21}^{k-1} \\ \mathbf{Y}_{21}^{l} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{11}^{k-1} \\ \mathbf{Y}_{11}^{l} \end{pmatrix} \right\|^2 > \left\| \begin{pmatrix} \mathbf{X}_{12}^{k-1} \\ \mathbf{Y}_{12}^{l} \end{pmatrix} - \begin{pmatrix} \mathbf{X}_{11}^{k-1} \\ \mathbf{Y}_{11}^{l} \end{pmatrix} \right\|^2 \Bigg| \boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \right] \end{aligned} \tag{5.3.9}$$

and integrating both side of (5.3.9) with respect to $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ we get $\Gamma(k-1+l-1) > \Gamma(k-1+l)$. This proves 5.3.6. Using similar argument and applying Proposition 5.3.1(a), (5.3.7) can be proved.

Lastly, If $b_k = 0$ for some k,

$$\begin{aligned}
\Gamma(k-1+l) =& \mathrm{P}\left[\left\|\begin{pmatrix}\mathbf{X}_{21}^{k-1}\\ \mathbf{Y}_{21}^{l}\end{pmatrix} - \begin{pmatrix}\mathbf{X}_{11}^{k-1}\\ \mathbf{Y}_{11}^{l}\end{pmatrix}\right\|^2 > \left\|\begin{pmatrix}\mathbf{X}_{12}^{k-1}\\ \mathbf{Y}_{12}^{l}\end{pmatrix} - \begin{pmatrix}\mathbf{X}_{11}^{k-1}\\ \mathbf{Y}_{11}^{l}\end{pmatrix}\right\|^2 \middle| \boldsymbol{\mu}_1, \boldsymbol{\mu}_2\right] \\
=& \mathrm{P}\left[U + V_l > 0\right] \quad \Big(\text{where } U = \left\|\mathbf{X}_{21}^{k-1} - \mathbf{X}_{11}^{k-1}\right\|^2 - \left\|\mathbf{X}_{12}^{k-1} - \mathbf{X}_{11}^{k-1}\right\|^2, \\
& \qquad\qquad \text{and } V_l = \left\|\mathbf{Y}_{21}^{l} - \mathbf{Y}_{11}^{l}\right\|^2 - \left\|\mathbf{Y}_{12}^{l} - \mathbf{Y}_{11}^{l}\right\|^2\Big) \\
=& \mathrm{P}\left[U + \sum_{i=1}^{l} W_i > 0\right] \qquad (\text{where } W_i\text{'s are i.i.d. with mean } 0) \\
=& \mathrm{P}\left[\frac{1}{\sqrt{l}}U + \frac{1}{\sqrt{l}}\sum_{i=1}^{l} W_i > 0\right] \to 1 - \Phi(0) = 1/2 \qquad (\text{by CLT}).
\end{aligned}$$

Hence, $\Gamma(p) \downarrow 1/2$ as $p \to \infty$. Using similar argument, it can be shown that $\Gamma'(p) \downarrow 1/2$ as $p \to \infty$. □

Thus, inclusion of nondiscriminating variables can only reduce the chance that the between-cluster is greater than the within-cluster distance. Also, if we go on including more and more nondiscriminating variables, the chance may be as bad as 1/2.

5.3.3 Threshold for Inclusion of a Variable

We have seen in the last section that nondiscriminating variables which have no mean separation across the different clusters (or zero between-clusters variance), should be dropped. Because of the continuity of the probability functions $\Gamma(d)$ and $\Gamma'(d)$ with respect to between-cluster variance of each component, it can be concluded that the contribution of a component with very small between-cluster variance should be small, and therefore such a component would not be worthy of inclusion. The question is: "how small is small?"

One may seek to answer this question by trying to include one variable at a time. This approach leads to another question: "is it possible to reach the optimal set of variables by including one variable at a time?" The following counter-example shows that the answer is 'no'.

Example Suppose that the within cluster dispersion matrix $\mathbf{W} = \mathbf{I}_{10}$ and

the between cluster dispersion matrix $\mathbf{B}$ is given by

$$\mathbf{B}_{10\times 10} = \begin{bmatrix} 10 & \mathbf{0} & 0 \\ \mathbf{0} & 0.3 \times \mathbf{I}_8 & \mathbf{0} \\ 0 & \mathbf{0} & 0.01 \end{bmatrix}.$$

Then, the values of $\Gamma(d)$ and $\Gamma^{'}(d)$ for $d = 1, 2, \ldots, 10$ are displayed in the following table.

d	$\Gamma(d)$	$\Gamma^{'}(d)$	d	$\Gamma(d)$	$\Gamma^{'}(d)$
1	0.8168	0.8116	6	0.8224	0.8117
2	0.8143	0.8096	7	0.8248	0.8137
3	0.8155	0.8088	8	0.8274	0.8155
4	0.8174	0.8094	9	0.8299	0.8176
5	0.8198	0.8103	10	0.8282	0.8145

Table 5.3.1: Probability of the event that the between-cluster Euclidean distance is greater than the within-cluster one for various subset sizes

So, if we would proceed sequentially, we would choose the first variable only. But actually, the optimal subset will consist of first 9 variables, as both $\Gamma(.)$ and $\Gamma^{'}(.)$ are maximum for $d = 9$.

The above example shows that we should not include the variables sequentially, as we may not reach the optimal subset of variables in that case.

5.3.4 Procedure for Selecting the Optimal Subset

As we cannot include one variable at a time, we need to compare the probabilities corresponding to the best subset of different subset sizes simultaneously in order to obtain the optimal subset. The procedure for choosing an optimal subset of a given size has already been outlined in Table 5.2.1. We only need to search for the best d with respect to the chosen criterion. One possible way to do so is to use simulation to estimate the probabilities $\Gamma(d)$ and $\Gamma'(d)$ for all values of d in a suitable range and choose the smaller of the two maximizers.

Since

$$\Gamma(d) = \mathrm{P}\left[\left\|\mathbf{X}_{11}^{d} - \mathbf{X}_{21}^{d}\right\|^2 > \left\|\mathbf{X}_{11}^{d} - \mathbf{X}_{12}^{d}\right\|^2\right],$$

and

$$\begin{pmatrix} \mathbf{X}_{11}^{d} - \mathbf{X}_{21}^{d} \\ \mathbf{X}_{11}^{d} - \mathbf{X}_{12}^{d} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} 2\mathbf{B}_d + 2\mathbf{I}_d & \mathbf{I}_d \\ \mathbf{I}_d & 2\mathbf{I}_d \end{pmatrix}\right], \tag{5.3.10}$$

where $\mathbf{B}_d$ is the leading principal submatrix of $\mathbf{B}$ of size $d \times d$, $\Gamma(d)$ can be well approximated by the average

$$\hat{\Gamma}(d) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\left(\|\mathbf{Y}_i\|^2 > \|\mathbf{Z}_i\|^2\right)},$$

where $(\mathbf{Y}_i, \mathbf{Z}_i)^T$ are i.i.d. from the normal distribution, mentioned in (5.3.10). In the same spirit, since

$$\Gamma^{'}(d) = \mathrm{P}\left[\left\|\mathbf{X}_{11}^d - \mathbf{X}_{21}^d\right\|^2 > \left\|\mathbf{X}_{12}^d - \mathbf{X}_{13}^d\right\|^2\right],$$

and

$$\begin{pmatrix} \mathbf{X}_{11}^d - \mathbf{X}_{21}^d \\ \mathbf{X}_{12}^d - \mathbf{X}_{13}^d \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} 2\mathbf{B}_d + 2\mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & 2\mathbf{I}_d \end{pmatrix}\right], \tag{5.3.11}$$

where $\mathbf{B}_d$ is as mentioned above, $\Gamma^{'}(d)$ can be well approximated by the average

$$\hat{\Gamma}^{'}(d) = n^{-1} \sum_{i=1}^{n} \mathbf{1}_{\left(\|\mathbf{Y}_i\|^2 > \|\mathbf{Z}_i\|^2\right)},$$

where $(\mathbf{Y}_i, \mathbf{Z}_i)^T$ are i.i.d. from the normal distribution, mentioned in (5.3.11).

5.4 Selection of Number of variables for Different Linkages

The criterion presented in the foregoing section allows us to select a d that would result in a probabilistically best segregated distance matrix. It is not clear however, whether a better segregated distance matrix would necessarily lead to better clustering for various types of linkage. In this section, we look for a method of choosing d that seeks to achieve better clustering by methods based on a particular type of linkage.

We consider the three standard types of linkage, namely single linkage, average linkage and the complete linkage.

5.4.1 Criterion Function

Suppose that there are two clusters C_1 and C_2. C_1 consists of n_1 observations $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ and C_2 consists of n_2 observations $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$.

A new observation, say $\mathbf{X}$ from C_1 will be associated to the first cluster if $D(\mathbf{X}, C_1) < D(\mathbf{X}, C_2)$, where

$$D(\mathbf{X}, C_i) = \begin{cases} D_s(\mathbf{X}, C_i) = \min\left\{\|\mathbf{X} - \mathbf{Y}_{ij}\|^2 : 1 \le j \le n_i\right\} \text{ for single linkage} \\ D_a(\mathbf{X}, C_i) = \left\|\mathbf{X} - n_i^{-1}\sum_{j=1}^{n_i}\mathbf{Y}_{ij}\right\|^2 \qquad \text{for average linkage} \\ D_c(\mathbf{X}, C_i) = \max\left\{\|\mathbf{X} - \mathbf{Y}_{ij}\|^2 : 1 \le j \le n_i\right\} \text{ for complete linkage.} \end{cases} \tag{5.4.1}$$

Following the argument given in Section 5.3.1, for any linkage, the subset of variables, for which the probability of the event $D(\mathbf{X}, C_1) < D(\mathbf{X}, C_2)$ is higher, is more amenable for correct clustering using that linkage method. So we take

$$\Gamma(d, n_1, n_2) = \begin{cases} \Gamma_s(d, n_1, n_2) = \mathrm{P}\left[D_s(\mathbf{X}^d, C_1) < D_s(\mathbf{X}^d, C_2)\right] \text{ for single linkage} \\ \Gamma_a(d, n_1, n_2) = \mathrm{P}\left[D_a(\mathbf{X}^d, C_1) < D_a(\mathbf{X}^d, C_2)\right] \text{ for average linkage} \\ \Gamma_c(d, n_1, n_2) = \mathrm{P}\left[D_c(\mathbf{X}^d, C_1) < D_c(\mathbf{X}^d, C_2)\right] \text{ for complete linkage,} \end{cases} \tag{5.4.2}$$

where $D(\mathbf{X}^d, C_i)$ is similarly defined as $D(\mathbf{X}, C_i)$ considering the first d many components only. We would like to maximize $\Gamma(d, n_1, n_2)$ with respect to d. However, the maximizer depends on n_1 and n_2. In order to avoid confusion, we would make the parsimonous choice of the smallest of all the d's that maximize $\Gamma(d, n_1, n_2)$ for different pairs of n_1 and n_2. This choice would ensure that the variables thus selected would have been included on the basis of $\Gamma(d, n_1, n_2)$ for *any* n_1 and n_2. We now present a limited simulation study in search of 'worst case' choices n_1^0 and n_2^2 such that

$$\arg\max_d \Gamma(d, n_1^0, n_2^0) \le \arg\max_d \Gamma(d, n_1, n_2) \ \forall\ n_1, n_2. \tag{5.4.3}$$

5.4.2 Simulation Plan

The simulation study consisted of two stages. In the first stage, we considered the identity matrix as the within-cluster dispersion matrix and three different between-cluster dispersion matrices, namely

$$\mathbf{B}_1 = \begin{bmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{bmatrix}, \mathbf{B}_2 = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & .5 \end{bmatrix}, \mathbf{B}_3 = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 \\ 0 & .8 & 0 & 0 & 0 \\ 0 & 0 & .8 & 0 & 0 \\ 0 & 0 & 0 & .8 & 0 \\ 0 & 0 & 0 & 0 & .8 \end{bmatrix}. \tag{5.4.4}$$

For each of the three linkage methods and three $\mathbf{B}_i$'s, we considered 9 pairs of values of n_1 and n_2. For each combination of the parameters, we tried to obtain one common threshold for the inclusion of 1, 3 or 5 additional variables. Larger the threshold, smaller is the optimal value of d for a given between-cluster dispersion matrix. Thus, while choosing n_1^0 and n_2^0, we would look for the largest threshold. The results are displayed in the following section.

From this study, we got some indication about n_1^0 and n_2^0 which will be reported later. In order to confirm that the pair (n_1^0, n_2^0) serves our purpose, we conducted the second stage of simulations. In this study, for each $\mathbf{B}_i$ we obtain d_{opt} by directly maximizing $\Gamma(d, n_1^0, n_2^0)$, and show that in each case $\Gamma(d, n_1, n_2) \leq \Gamma(d_{\text{opt}}, n_1, n_2)$, for all $d < d_{\text{opt}}$ for three different pairs of (n_1, n_2), namely $n_1 < n_2, n_1 = n_2$ and $n_1 > n_2$. This is displayed in the Figures 5.4.3, 5.4.2 and 5.4.3, respectively.

5.4.3 Results and Discussion

In the following tables, we present the common threshold values for the inclusion of r many variables in the clustering procedure using the three standard linkage methods, when the cluster sizes are n_1 and n_2 respectively and $\mathbf{B}$ is the between-cluster dispersion matrix before including any variable.

Table 5.4.1 shows that if average linkage is used, the threshold is maximum when n_1 is smallest and n_2 is largest. Similar phenomenon is observed for single linkage, as shown in Table 5.4.2. However, in case of complete linkage, the threshold is maximized when n_1 is largest and n_2 is smallest, as shown in Table 5.4.3. This gives an indication that in the case of average linkage and single linkage, the threshold will maximize if $n_1 \downarrow$ and $n_2 \uparrow$ whereas, in the case of complete linkage the threshold maximizes as $n_2 \downarrow$ and $n_2 \uparrow$.

The simulations indicate that for single and average linkages, the worst case combination may be $n_1^0 = 1$, $n_2^0 = \infty$, while for complete linkage, the worst case combination may be $n_1^0 = \infty$, $n_2^0 = 1$. In order to verify this, we assumed n_1^0 and n_2^0 to be as above and tried to see whether (5.4.3) holds. Figure 5.4.1 shows that for average linkage the condition (5.4.3) seems to hold.

B	Number of Variables to be added											
	1				3				5			
B1	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.5	1.05	1.15	1	.47	.93	.97	1	.43	.92	0.933
	50	.01	.05	.1	50	0	0.005	.03	50	0	.02	.03
	100	0	.01	.03	100	0	.005	.02	100	0	.02	.02
B2	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.5	.5	.5	1	.45	.5	.5	1	.5	.5	.5
	50	0	0	0	50	.001	.05	.05	50	0	.003	.003
	100	0	0	0	100	0	0	.001	100	0	.003	.003
B3	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.32	.71	.74	1	0.296	0.696	0.704	1	.27	.69	.69
	50	0	.02	.04	50	0	0.024	0.032	50	0	.02	.02
	100	0	.001	.001	100	0	0	0.016	100	0	0	.02

Table 5.4.1: Threshold values for different combinations $\mathbf{B}, n_1, n_2$ and r, the number of variables to be added, in case of average linkage

B	Number of Variables to be added											
	1				3				5			
B1	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.55	.75	.95	1	.45	.70	.80	1	.2	.70	.75
	50	.05	.30	.10	50	.05	.25	.25	50	0	.15	.15
	100	0	0	.25	100	0	.10	.25	100	0	.15	.10
B2	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.21	.50	.63	1	.16	.49	.56	1	.02	.49	.54
	50	0	.05	.06	50	0	.005	.01	50	0	.005	.005
	100	0	.005	.07	100	0	.005	.02	100	0	0	.005
B3	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.26	.33	.46	1	.15	.32	.35	1	.06	.28	.32
	50	.008	.07	.09	50	.008	.03	.08	50	.008	.04	.08
	100	0	.008	.024	100	0	0	.008	100	0	.008	.01

Table 5.4.2: Threshold values for different combinations $\mathbf{B}, n_1, n_2$ and r, the number of variables to be added, in case of single linkage

B	Number of Variables to be added											
	1				3				5			
B1	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.25	.1	.05	1	.2	.1	0	1	.15	.05	0
	50	.551	.2	.05	50	.45	.2	.005	50	.25	.05	.005
	100	.65	.2	.1	100	.5	.25	.025	100	.5	.2	.2
B2	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.05	.005	0	1	.045	0	0	1	.02	0	0
	50	.33	.15	.005	50	.05	.05	0	50	.05	.025	0
	100	.39	.15	.02	100	.35	.02	.005	100	.1	.02	0
B3	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100	n_1 \ n_2	1	50	100
	1	.16	.016	.005	1	.07	.01	0.005	1	.04	.005	0
	50	.34	.02	0	50	.18	.02	.008	50	.09	.02	.005
	100	.37	.164	.032	100	.29	.14	.02	100	.11	.05	.002

Table 5.4.3: Threshold values for different combinations $\mathbf{B}, n_1, n_2$ and r, the number of variables to be added, in case of complete linkage

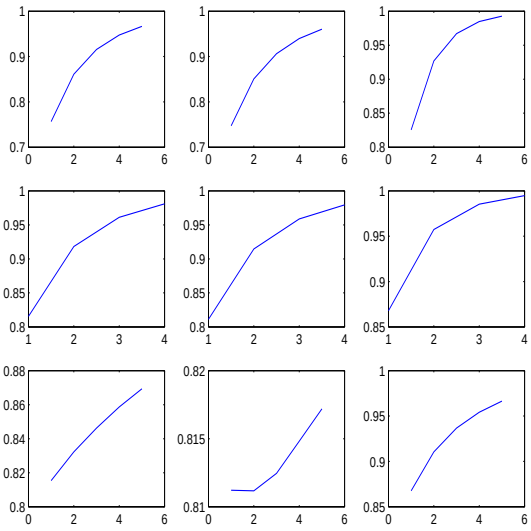


Figure 5.4.1: Plot of $\Gamma_a(d, n_1, n_2)$ against dimension d. The columns of the subplots correspond to the cases $(n_1, n_2) = (1, 1), (1, 50), (50, 1)$ respectively, whereas the rows correspond to $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$, respectively.

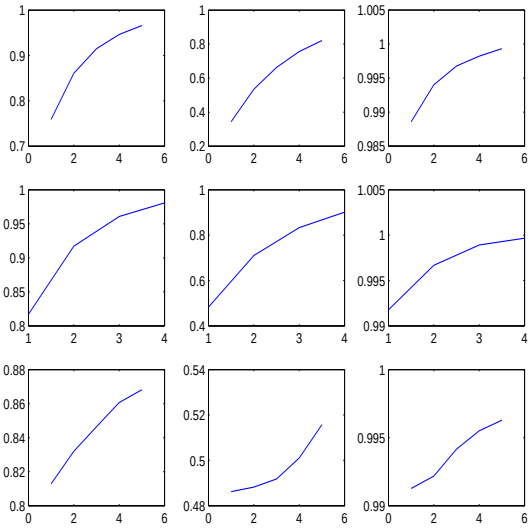


Figure 5.4.2: Plot of $\Gamma_s(d, n_1, n_2)$ against dimension d. The columns of the subplots correspond to the cases $(n_1, n_2) = (1, 1), (1, 50), (50, 1)$ respectively, whereas the rows correspond to $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$, respectively.

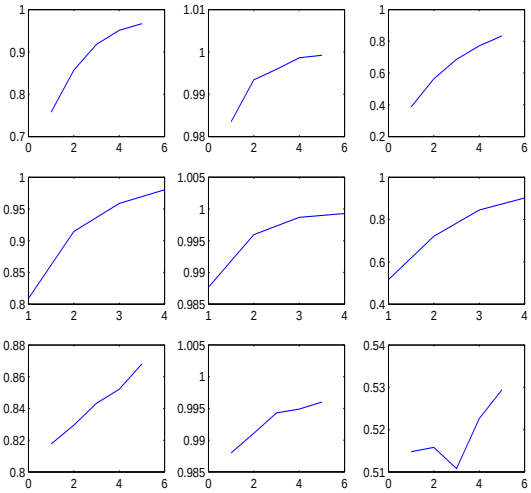


Figure 5.4.3: Plot of $\Gamma_c(d, n_1, n_2)$ against dimension d. The columns of the subplots correspond to the cases $(n_1, n_2) = (1, 1), (1, 50), (50, 1)$ respectively, whereas the rows correspond to $\mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3$, respectively.

5.5 Conclusion

From the simulation study, it seems that the conservative choice of n_1 and n_2 for the three linkage methods are as given in Table 5.5.1. If these conjectures hold, we need to consider only the pair (n_1^0, n_2^0) and maximize $\Gamma(d, n_1^0, n_2^0)$ with respect to d, as the corresponding d_{opt} would be a conservative choice for any other (n_1, n_2). Simulation plan for maximizing $\Gamma(d, n_1^0, n_2^0)$ would be much simpler than searching over all n_1 and n_2 and the corresponding optimal d.

Linkage	n_1^0	n_2^0
Single	1	∞
Average	1	∞
Complete	∞	1

Table 5.5.1: Conservative choice of n_1 and n_2 for different types of linkage

Chapter 6

Clustering with or without Training Data

In Chapter 5, we had assumed both $\mathbf{W}$ and $\mathbf{B}$ to be known and discussed the ill effects of nondiscriminating variables in clustering. When these matrices are unknown, their estimation itself may become a difficult task because of high dimensionality. In this chapter, we discuss these aspects of clustering.

In usual clustering problems, we do not have any knowledge of$\mathbf{W}$ and $\mathbf{B}$ separately. So the techniques for known $\mathbf{W}$ and $\mathbf{B}$ cannot be applied. We can only hope that $\mathbf{B}$ is sufficiently 'larger' than $\mathbf{W}$ in some sense so that the correct clusters would be recognizable from the data at hand. We formalize this notion in the next section, and examine conditions under which reasonable clustering would indeed be possible.

In some cases, clustering of the main data set ('test' data) is preceded by analysis of a typically smaller data set where the cluster memberships are known ('training data'). We can estimate $\mathbf{W}$ and $\mathbf{B}$ from the training data, but because of high dimensionality or shortage of data, the estimates may not be reliable. If these estimates are substituted in place of $\mathbf{W}$ and $\mathbf{B}$, the resulting 'best' linear combinations may differ from the optimal ones substantially. We discuss some regularization techniques to overcome this difficulty, and examine their effectiveness through simulation.

6.1 Clustering with No Training Data

In this case, we don't have any prior knowledge of the between-cluster ($\mathbf{B}$) and within-cluster ($\mathbf{W}$) variance-covariance matrices. So, we cannot estimate the eigenvectors of $\mathbf{W}^{-\frac{1}{2}}\mathbf{B}\mathbf{W}^{-\frac{1}{2}}$ corresponding to its largest eigenval-

ues, which give the optimal linear combinations according to our criterion. However, we can estimate the total dispersion matrix $\mathbf{W}+\mathbf{B}$. We may choose the eigenvectors of the estimate of $\mathbf{W}+\mathbf{B}$ corresponding to its largest eigenvalues, as the discriminating directions. In this section, we study the performance of these linear combinations in terms of our criterion. For the time, being we study the performance of these eigenvectors, disregarding the fact that the latter matrix is estimated from the data. We denote by $\mathbf{L}_1$ the semi-orthogonal $p \times d$ matrix whose columns are the eigenvectors of $\mathbf{W}+\mathbf{B}$ corresponding to its largest d many eigenvalues.

The eigenvectors of $\mathbf{W}+\mathbf{B}$ are expected to perform well, intuitively, if in some directions the between-cluster variance is very high compared to the of within-cluster variance. For example, if the larger eigenvalues of $\mathbf{B}$ are very high compared to those of $\mathbf{W}$, the eigenvectors of $\mathbf{W}+\mathbf{B}$ corresponding to its large eigenvalues are quite likely to be close to the optimal directions, and hence have high discriminating power. To study the performance of $\mathbf{L}_1$, we consider the following norm on n.n.d. matrices.

$$\|\mathbf{A}\|_d = \text{sum of } d \text{ largest eigenvalues of } \mathbf{A}.$$

That $\|.\|_d$ is a norm follows from the fact that by Corollary 5.2.3,

$$\|\mathbf{A}\|_d = \max_{\{\mathbf{L}_{p\times d}:\mathbf{L}^T\mathbf{L}=\mathbf{I}_d\}} \text{trace}(\mathbf{L}^T\mathbf{A}\mathbf{L}) = \max_{\mathbf{L}\in\mathbb{L}} \text{trace}\left[(\mathbf{L}^T\mathbf{A}\mathbf{L})(\mathbf{L}^T\mathbf{L})^{-1}\right], \tag{6.1.1}$$

where $\mathbb{L} = \{\mathbf{L} : \mathbf{L} \text{ is } p\times d\}$, as defined in (5.2.3). This norm has an important property, namely

Lemma 6.1.1.

$$\left\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\right\|_d \le \|\mathbf{A}\|_d\|\mathbf{B}\|_d.$$

Proof. We observe that

$$\begin{aligned}
\left\|\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\right\|_d &= \max_{\mathbf{L}\in\mathbb{L}} \text{trace}\left[\left(\mathbf{L}^T\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\mathbf{L}\right)\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\right] \\
&= \max_{\mathbf{L}\in\mathbb{L}} \text{trace}\left[\left(\mathbf{L}^T\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\mathbf{L}\right)\left(\mathbf{L}^T\mathbf{A}\mathbf{L}\right)^{-1}\left(\mathbf{L}^T\mathbf{A}\mathbf{L}\right)\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\right] \\
&\le \max_{\mathbf{L}\in\mathbb{L}} \text{trace}\left[\left(\mathbf{L}^T\mathbf{A}^{\frac{1}{2}}\mathbf{B}\mathbf{A}^{\frac{1}{2}}\mathbf{L}\right)\left(\mathbf{L}^T\mathbf{A}\mathbf{L}\right)^{-1}\right] \\
&\qquad \times \max_{\mathbf{L}\in\mathbb{L}} \text{trace}\left[\left(\mathbf{L}^T\mathbf{A}\mathbf{L}\right)\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\right] \\
&= \max_{\mathbf{M}\in\mathbb{L}} \text{trace}\left[\left(\mathbf{M}^T\mathbf{B}\mathbf{M}\right)\left(\mathbf{M}^T\mathbf{M}\right)^{-1}\right]\|\mathbf{A}\|_d \quad (\text{putting } \mathbf{A}^{\frac{1}{2}}\mathbf{L} = \mathbf{M}) \\
&= \|\mathbf{A}\|_d\|\mathbf{B}\|_d
\end{aligned} \tag{6.1.2}$$

Hence, the inequality holds. □

Next, we obtain a lower bound for $\Psi(\mathbf{L}_1)$, where $\Psi(\cdot)$ is the criterion function defined in (5.2.5).

Proposition 6.1.2.

$$\Psi(\mathbf{L}_1) \geq \frac{\|\mathbf{W}+\mathbf{B}\|_d}{\|\mathbf{W}\|_d} - d$$

Proof. By the choice of $\mathbf{L}_1$,

$$\mathbf{L}_1^T(\mathbf{W}+\mathbf{B})\mathbf{L}_1 = \text{Diag}(\lambda_1, \ldots, \lambda_d) = \mathbf{D} \text{ say,}$$

where $\lambda_1, \ldots, \lambda_d$ are the d largest eigenvalues of $\mathbf{W}+\mathbf{B}$. So,

$$\begin{aligned}
\Psi(\mathbf{L}_1) + d &= \text{trace}\left[\left(\mathbf{L}_1^T\mathbf{B}\mathbf{L}_1\right)\left(\mathbf{L}_1^T\mathbf{W}\mathbf{L}_1\right)^{-1}\right] + d \\
&= \text{trace}\left[\mathbf{L}_1^T\left(\mathbf{W}+\mathbf{B}\right)\mathbf{L}_1\left(\mathbf{L}_1^T\mathbf{W}\mathbf{L}_1\right)^{-1}\right] \\
&= \text{trace}\left[\mathbf{D}\left(\mathbf{L}_1^T\mathbf{W}\mathbf{L}_1\right)^{-1}\right] \\
&= \sum_{i=1}^{d} \lambda_i \left(\mathbf{L}_1^T\mathbf{W}\mathbf{L}_1\right)^{ii} \\
&\geq \sum_{i=1}^{d} \frac{\lambda_i}{(\mathbf{L}_1^T\mathbf{W}\mathbf{L}_1)_{ii}} \\
&\geq \frac{\sum_{i=1}^{d}\lambda_i}{\sum_{i=1}^{d}(\mathbf{L}_1^T\mathbf{W}\mathbf{L}_1)_{ii}} \\
&= \frac{\|\mathbf{W}+\mathbf{B}\|_d}{\text{trace}(\mathbf{L}_1^T\mathbf{W}\mathbf{L}_1)} \\
&\geq \frac{\|\mathbf{W}+\mathbf{B}\|_d}{\|\mathbf{W}\|_d} \qquad (\text{as } \mathbf{L}_1^T\mathbf{L}_1 = \mathbf{I}_d). \qquad (6.1.3)
\end{aligned}$$

Hence the inequality is established. □

Using Lemma 6.1.1 and Proposition 5.2.1, we can compare the performance of $\mathbf{L}_1$ with that of the optimal linear combinations $\mathbf{L}_0$ that maximizes $\Psi(\cdot)$.

Proposition 6.1.3. *If* $\mathbf{L}_0$ *maximizes* $\Psi(\cdot)$*, then*

$$1 \geq \frac{\Psi(\mathbf{L}_1)+d}{\Psi(\mathbf{L}_0)+d} \geq \frac{1}{\|\mathbf{W}\|_d\|\mathbf{W}^{-1}\|_d}.$$

Proof. As $\mathbf{L}_0$ maximizes $\Psi(\cdot)$,

$$\begin{aligned}\Psi(\mathbf{L}_0)+d =& \max_{\mathbf{L}\in\mathbb{L}} \text{ trace}\left[\left(\mathbf{L}^T\mathbf{B}\mathbf{L}\right)\left(\mathbf{L}^T\mathbf{W}\mathbf{L}\right)^{-1}\right]+d \\ =& \max_{\mathbf{L}\in\mathbb{L}} \text{ trace}\left[\mathbf{L}^T\left(\mathbf{W}+\mathbf{B}\right)\mathbf{L}\left(\mathbf{L}^T\mathbf{W}\mathbf{L}\right)^{-1}\right] \\ =& \left\|\mathbf{W}^{-\frac{1}{2}}(\mathbf{W}+\mathbf{B})\mathbf{W}^{-\frac{1}{2}}\right\|_d \\ \leq& \left\|\mathbf{W}^{-1}\right\|_d\|\mathbf{W}+\mathbf{B}\|_d \qquad \text{(using lemma 6.1.1).} \end{aligned} \tag{6.1.4}$$

Combining this result with that of proposition 6.1.2, we get

$$\Psi(\mathbf{L}_0)+d \geq \Psi(\mathbf{L}_1)+d \geq \frac{\|\mathbf{W}+\mathbf{B}\|_d}{\|\mathbf{W}\|_d} \geq \frac{\Psi(\mathbf{L}_0)+d}{\|\mathbf{W}\|_d\|\mathbf{W}^{-1}\|_d}.$$

Thus, the inequality is established. □

The above lower bound is attainable. For example, if $\mathbf{B}$ and $\mathbf{W}$ are both diagonal, say $\mathbf{B} = \text{Diag}(\lambda_1,\ldots,\lambda_p)$ and $\mathbf{W} = \text{Diag}(\nu_1,\ldots,\nu_p)$, with $\lambda_1 = \lambda_2 > \lambda_3 = \cdots = \lambda_p$ and $\nu_1 > \nu_2 = \cdots = \nu_p$. If we consider the case $d = 1$, $\Psi(\cdot)$ is maximized at $\mathbf{L}_0 = (0,1,0,\ldots,0)^T$. However, if we use $\mathbf{W}+\mathbf{B}$, the largest eigenvector is $\mathbf{L}_1 = (1,0,\ldots,0)^T$. So,

$$\frac{\Psi(\mathbf{L}_1)+1}{\Psi(\mathbf{L}_0)+1} = \frac{\lambda_1/\nu_1}{\lambda_2/\nu_2} = \frac{\nu_2}{\nu_1} = \frac{\lambda_{\min}(\mathbf{W})}{\lambda_{\max}(\mathbf{W})} = \frac{1}{\|\mathbf{W}\|_1\|\mathbf{W}^{-1}\|_1}.$$

The upper bound is also sharp, as one can easily verify by considering $\mathbf{B} = \text{Diag}(\lambda_1,\ldots,\lambda_p)$, $\mathbf{W} = \text{Diag}(\nu_1,\ldots,\nu_p)$, with $\lambda_1 > \lambda_2 = \cdots = \lambda_p$, $\nu_1 > \nu_2 = \cdots = \nu_p$ and $\lambda_1/\nu_1 > \lambda_2/\nu_2 = \cdots = \lambda_p/\nu_p$.

The lower bound depends on the eigenvalues of $\mathbf{W}$. If the eigenvalues have very high ratios among them, the lower bound can be very small. For $d = 1$, the lower bound is nothing but $\frac{\lambda_{\min}(\mathbf{W})}{\lambda_{\max}(\mathbf{W})}$, the inverse of the condition number of $\mathbf{W}$, which can sometimes be very small.

However, if the large eigenvalues of $\mathbf{B}$ are very high compared to those of $\mathbf{W}$, i.e. $\|B\|_d \gg \|W\|_d$, then from proposition 6.1.2, it follows that

$$\Psi(\mathbf{L}_1)+d \geq \frac{\|\mathbf{W}+\mathbf{B}\|_d}{\|\mathbf{W}\|_d} \geq \frac{\|\mathbf{B}\|_d}{\|\mathbf{W}\|_d} \gg 1.$$

Thus, the eigenvectors of $\mathbf{W}+\mathbf{B}$ can perform reasonably well.

In the following subsection, we have a practical example, where the eigenvectors of $\mathbf{W}+\mathbf{B}$ work very well as discriminating directions.

6.2 An Example: Clustering of DNA Sequences

6.2.1 Data Description

The data which we considered consists of tetra-nucleotide frequencies derived from 16S and 18S ribosomal DNA sequences from 24 organisms (6 bacteria, 6 archaea, 6 fungi and 6 gymnosperm plants - 16S for bacteria and archaea while 18S for fungi and gymnosperms). For each organism, there are 256 variables, each variable being the percentage of occurrence of a particular tetra-nucleotide in a ribosomal DNA sequence. Tetra-nucleotides are 4-tuples formed by characters A, T, C and G, – hence there are 256 of them.

6.2.2 Methodology Used

One expects that 16S and 18S ribosomal DNA sequences carry some signatures of the species, and these can be reflected in the tetra-nucleotide frequencies. So one hopes that clustering based on these variables or variables derived from them may lead to groups that are biologically homogeneous.

Since sum of all the variables is 100, the first 255 variables were considered for clustering. So, in this case $p = 255$ and $N = 24$. The sample dispersion matrix $\mathbf{S}$ for all the observations was obtained and its principal components were considered. For different values of d, the leading d many principal components of $\mathbf{S}$ were used to transform the data to a d-dimensional observation. Then average linkage clustering was used to find the groups.

In the Table 6.2.1, we display the clusters obtained d many leading principal components, for $d = 1, 3, 5$.In that matrix, M represents the heterogeneity matrix, namely M_{ij} is the number o f common elements in the true i^{th} group and the j^{th} cluster obtained using average linkage clustering method.

6.2.3 Results and Discussion

This is a high dimensional clustering problem, as $p = 255$ is very large compared to $N = 24$. If we use two leading principal components of the sample dispersion matrix as the discriminating directions, the resulting clusters are quite close to what is expected from the biological point of view. However, if the number of linear combinations are increased, the performance become worse. Table 6.2.1 shows that the clusters which we obtain by using 5 linear combinations, differ from the true one considerably. This shows that if the number of linear combinations is increased, it may not always improve the

clustering.

d	Cluster1		Cluster2		Cluster3		Cluster4		M			
	1	2	4		5	15	13	14	4	1	1	0
	3	6			17	19	16	18	6	0	0	0
1	7	8			20	21			0	0	2	4
	9	10			22	23			0	0	6	0
	11	12			24							
	1	2	6	7	13	14	15	17	5	1	0	0
2	3	4	8	9	16	18	19	20	0	6	0	0
	5		10	11			21	22	0	0	4	2
			12				23	24	0	0	0	6
	1	2	4	5	7	8	13	14	4	2	0	0
5	3	6			9	10	15	16	0	0	6	0
					11	12	17	18	0	0	0	6
							19	20	0	0	0	6
							21	22				
							23	24				

Table 6.2.1: Results of clustering using d many principal components. Here M is the heterogeneity matrix.

6.3 Clustering with Training Data

In this case, we can get estimates of $\mathbf{B}$ and $\mathbf{W}$, say $\hat{\mathbf{B}}$ and $\hat{\mathbf{W}}$, respectively, using training samples. Given d (the number of linear combinations to be chosen), we can maximize trace$[(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})\ (\mathbf{L}^T\hat{\mathbf{W}}\mathbf{L})^{-1}]$ with respect to $\mathbf{L} \in \mathbb{L}$, in order to estimate the true optimizer of our criteria function. However, if the dimension p of the observations is very high compared to the number of observations, the estimates of $\mathbf{B}$ and $\mathbf{W}$ become poor and unstable. The linear combinations obtained by using these unreliable estimates, are often substantially different from the optimal linear combinations. In this situation, regularization techniques may be quite helpful in reducing the instability of the estimates and hence improve the performance of the corresponding linear combinations. In this study, we have considered three regularization techniques. These are discussed in the next subsections.

6.3.1 Low Rank Approximation

Generally, when dimension of the observations is very high in relation to the sample size, the small eigenvalues of the covariance matrix are underestimated. Consrquently, the associated eigenvectors are estimated with a very high variance. In the present problem, eigenvectors of $\mathbf{W}$ corresponding to its small eigenvalues are poorly estimated. This may cause the resulting linear combinations of variables to perform poorly.

One possible way to cope with this problem is to consider a low rank approximation of $\hat{\mathbf{W}}$. In this regularization, small eigenvalues of $\hat{\mathbf{W}}$ are set to 0, and the search for the optimal $\mathbf{L}$ is restricted to the eigenspace of $\hat{\mathbf{W}}$ corresponding to its nonzero eigenvalues. Thus, we can neutralize the effect of the poorly estimated and unstable eigenvectors of $\mathbf{W}$, at the cost of confining to a smaller subspace of $\mathbb{R}^p$. So, this regularization is expected to perform well, if sufficient separation is present even in the smaller subspace, corresponding to the large eigenvalues of $\mathbf{W}$. To be more precise, let us consider the spectral decomposition of $\hat{\mathbf{W}}$, $\hat{\mathbf{W}} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$, where $\mathbf{\Lambda}$ =Diag$(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Suppose that $\mathbf{\Lambda}_1$ =Diag$(\lambda_1, \ldots, \lambda_k)$ contains the large eigenvalues and $\mathbf{\Lambda}_2$ =Diag$(\lambda_{k+1}, \ldots, \lambda_p)$ contains the small eigenvalues of $\hat{\mathbf{W}}$. So, we can write

$$\hat{\mathbf{W}} = \mathbf{P}_1\mathbf{\Lambda}_1\mathbf{P}_1^T + \mathbf{P}_2\mathbf{\Lambda}_2\mathbf{P}_2^T = \hat{\mathbf{W}}_1 + \hat{\mathbf{W}}_2.$$

In low rank approximation, we ignore the second term and maximize the trace of $[(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}_1\mathbf{L})^{-1}]$ instead of the trace of $[(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}\mathbf{L})^{-1}]$ with respect to $\mathbf{L}$, $C(\mathbf{L}) \subseteq C(\hat{\mathbf{W}}_1)$. Now, if $C(\mathbf{L}) \subseteq C(\hat{\mathbf{W}}_1)$, then $\mathbf{L}^T\hat{\mathbf{W}}\mathbf{L} =$

$\mathbf{L}^T\hat{\mathbf{W}}_1\mathbf{L}$. So, in low rank approximation, we maximize the same quantity, trace of $[(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}\mathbf{L})^{-1}]$, but in a lower dimensional space, namely within $C(\hat{\mathbf{W}}_1)$, which leads to more reliable estimate of the corresponding population analogue. So, the strategy would work well if

$$\max_{\mathbf{L}\,:\,C(\mathbf{L})\subseteq C(\mathbf{W}_1)} \text{trace}\left[\left(\mathbf{L}^T\mathbf{B}\mathbf{L}\right)\left(\mathbf{L}^T\mathbf{W}\mathbf{L}\right)^{-1}\right]$$

is large.

6.3.2 Shrinkage Towards Diagonal Matrix

Instead of using a low rank approximation, we can regularize $\hat{\mathbf{W}}$ by shrinking it towards $\hat{\mathbf{D}}$ =Diag($\hat{\mathbf{W}}$), i.e. a diagonal matrix having the same diagonal entries as of $\hat{\mathbf{W}}$. To be more precise, we consider a convex combination of $\hat{\mathbf{W}}$ and $\hat{\mathbf{D}}$, namely

$$\hat{\mathbf{W}}(\lambda) = (1-\lambda)\hat{\mathbf{W}} + \lambda\hat{\mathbf{D}},$$

and maximize the trace of $[(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}(\lambda)\mathbf{L})^{-1}]$ with respect to $\mathbf{L}$ in order to estimate the best discriminating directions. Due to high dimensionality, if the estimate $\hat{\mathbf{W}}$ becomes unstable, shrinkage towards diagonal matrix using small λ is expected to improve the estimate by reducing the eigenvalue distortion and thereby help to obtain better discriminating directions. In practice, the shrinkage parameter λ can be estimated by considering a grid of $[0,1]$ and choosing the grid point for which $\max_{\mathbf{L}}$ trace $\left[(\mathbf{L}^T\mathbf{B}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}(\lambda)\mathbf{L})^{-1}\right]$ is maximum.

6.3.3 Regularizing Moore-Penrose G-Inverse of $\hat{\mathbf{W}}$

In order to maximize trace of $(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}\mathbf{L})^{-1}$ with respect to $\mathbf{L}$, $C(\mathbf{L}) \subseteq C(\hat{\mathbf{W}})$, we need to consider the eigenvector of $\hat{\mathbf{W}}^{-\frac{1}{2}}\mathbf{B}\hat{\mathbf{W}}^{-\frac{1}{2}}$ where $\hat{\mathbf{W}}^{-\frac{1}{2}}$ is the square root of $\hat{\mathbf{W}}^{+}$ (the Moore-Penrose g-inverse of $\hat{\mathbf{W}}$), corresponding to its largest eigenvalue and transform it suitably. Since the contribution of $\hat{\mathbf{W}}$ is through $\hat{\mathbf{W}}^{+}$, regularizing $\hat{\mathbf{W}}^{+}$ itself instead of $\hat{\mathbf{W}}$ may improve the performance of the resulting linear combinations. On the basis of this intuitive reasoning, we considered the regularization of $\hat{\mathbf{W}}^{+}$. We take a convex combination of $\hat{\mathbf{W}}^{+}$ and $\hat{\mathbf{D}}_*$ =Diag($\hat{\mathbf{W}}^{+}$) and then use it in place of $\hat{\mathbf{W}}^{+}$ to get the optimal linear combinations. So, the regularization is given by

$$\hat{\mathbf{W}}_*(\lambda) = \left[(1-\lambda)\hat{\mathbf{W}}^{+} + \lambda\hat{\mathbf{D}}_*\right]^{+}.$$

This regularization is expected to improve $\hat{\mathbf{W}}^+$, by reducing its instability due to high dimension and thus improve the resulting linear combinations. As in the previous case, we consider a grid of $[0, 1]$ and estimate the shrinkage parameter by the grid point for which $\max_{\mathbf{L}}$ trace $\left[(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}\mathbf{L})^{-1}\right]$ is maximum.

6.4 Simulation Study

In order to compare the regularization methods, we considered a 50-dimensional Gaussian population consisting of 5 clusters, each containing 10 observations. Three different choices of $\mathbf{W}$ and $\mathbf{B}$ were considered in the following three examples. We obtained $\hat{\mathbf{W}}$ from the simulated data and regularized it using all the three regularization methods, discussed above. For each of the regularized $\hat{\mathbf{W}}$, we obtained the maximizer of $\left[(\mathbf{L}^T\hat{\mathbf{B}}\mathbf{L})(\mathbf{L}^T\hat{\mathbf{W}}\mathbf{L})^{-1}\right]$ and evaluated its performance by calculating $\left[(\mathbf{L}^T\mathbf{B}\mathbf{L})(\mathbf{L}^T\mathbf{W}\mathbf{L})^{-1}\right]$ for each of them.

In the figures, we have plotted the performance of the optimizing linear combinations for different extents of regularization.

• The blue line represents low rank approximation. It is plotted against the rank of $\hat{\mathbf{W}}$, which decreases as we move from left to right.
• The red line represents the performance of "shrinkage towards diagonal matrix", and it is plotted against λ, the shrinkage parameter. As we move from left to right, we go away from $\hat{\mathbf{W}}$ towards its diagonal.
• The black line represents the performance of "regularization of $\hat{\mathbf{W}}^+$". This is also plotted against the shrinkage parameter λ. As we move from left to right, we go away from $\hat{\mathbf{W}}$.
All three graphs are plotted in the same vertical scale to facilitate comparison. Horizontal scales are not comparable.

6.4.1 Results of Simulation Study

Example 1 In the first example, the within-cluster variation matrix $\mathbf{W}$ was chosen to have very high ratio of the large and low eigenvalues. The small eigenvalues were of the order 10^{-3}, whereas the large eigenvalues were of the order 10^2. More precisely, the eigenvalues were $1 \times 10^{-3}, 2 \times 10^{-3}, \ldots, 25 \times 10^{-3}, 1 \times 10^2, 2 \times 10^2, \ldots, 25 \times 10^2$. So, the condition number of $\mathbf{W}$ was 25×10^5. The eigenvectors were chosen randomly. The between-cluster matrix $\mathbf{B}$ was chosen to be a diagonal matrix, with all the diagonal entries, except the last five, as 10^{-3}. The last five diagonal entries were chosen to be 10^4, to make sure that only the last five components have enough discriminating power.

The performance of different regularization methods are displayed in the following figure.

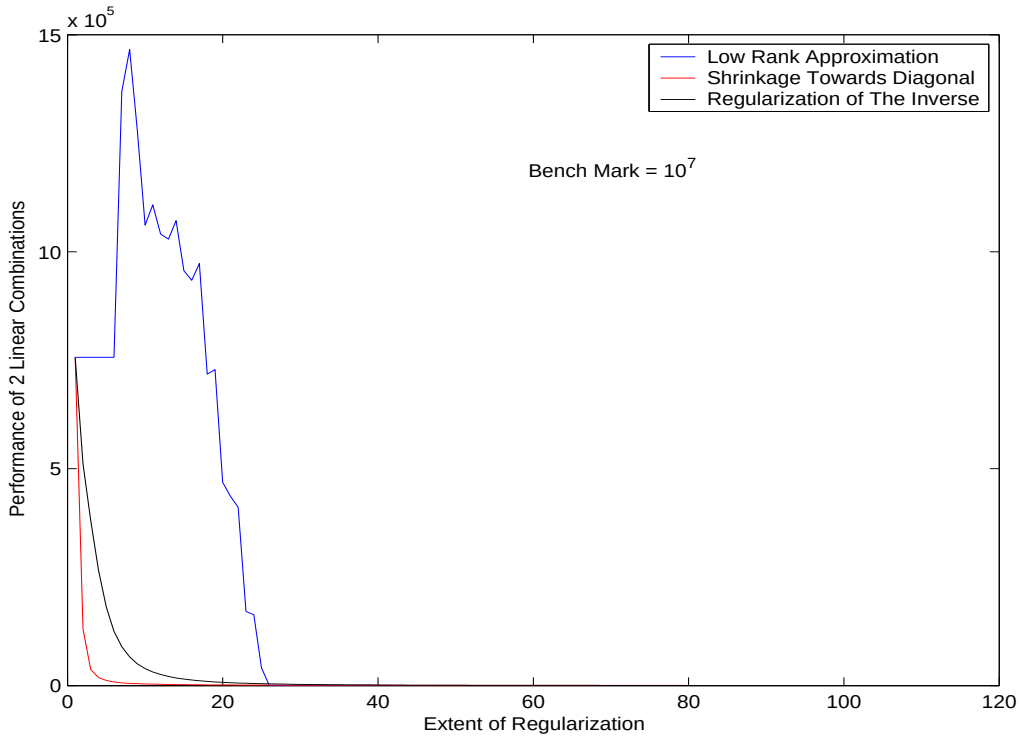


Figure 6.4.1: Performance of various regularization methods for Example 1

Example 2 In the second example, the within-cluster variation matrix $\mathbf{W}$ was chosen to have moderately high ratio among the large and low eigenvalues. The small eigenvalues were of the order 10^{-3}, whereas the large eigenvalues were between 1 and 25. More precisely, the eigenvalues were $1 \times 10^{-3}, 2 \times 10^{-3}, \ldots, 25 \times 10^{-3}, 1, 2, \ldots, 25$. So, the condition number of $\mathbf{W}$ was 25×10^3. The eigenvectors were chosen randomly. The between-cluster matrix $\mathbf{B}$ was chosen to be a diagonal matrix, with all the diagonal entries, except the last five, as 10^{-3}. The last five diagonal entries were chosen to be 10^2 to make sure that only the last five components have enough discriminating power.

The performance of different regularization methods are displayed in the following figure.

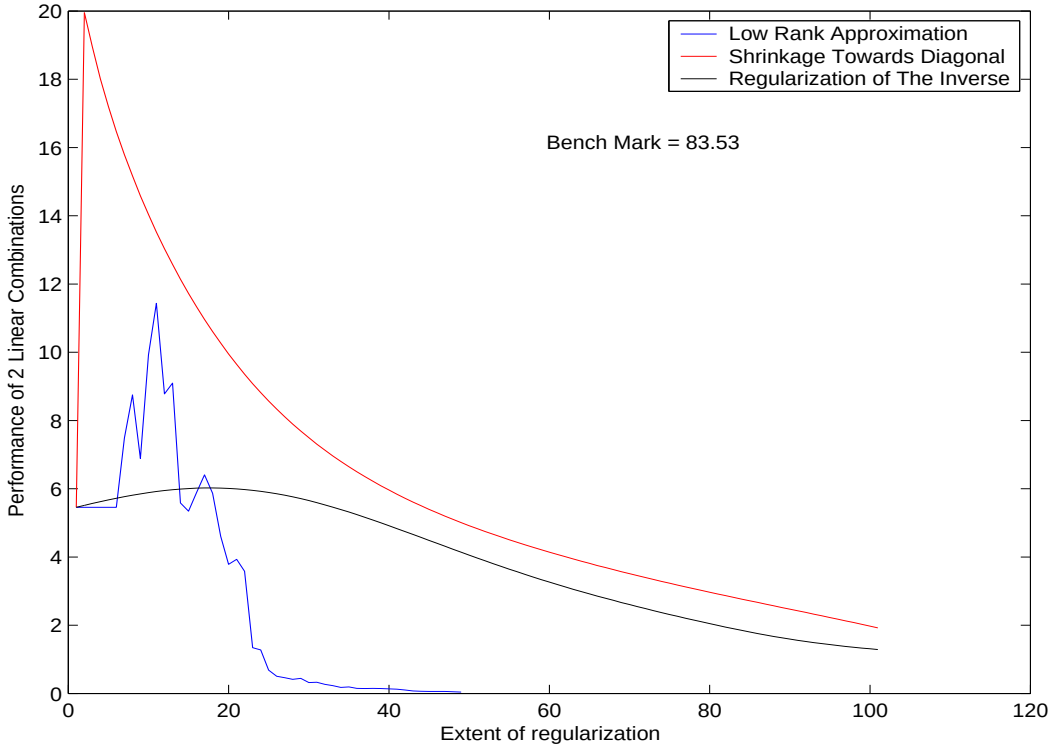


Figure 6.4.2: Performance of various regularization methods for Example 2

Example 3 In the third example, the within-cluster variation matrix **W** was chosen to have low ratio among the large and low eigenvalues. The eigenvalues were $1, 2, \ldots, 50$. So, the condition number of **W** was 50. The eigenvectors were chosen randomly. The between-cluster matrix **B** was chosen to be a diagonal matrix, with all the diagonal entries, except the last five, as 10^{-3}. The last five diagonal entries were chosen to be 10^4, to make sure that only the last five components have enough discriminating power.

The performance of different regularization methods are displayed in the following figure.

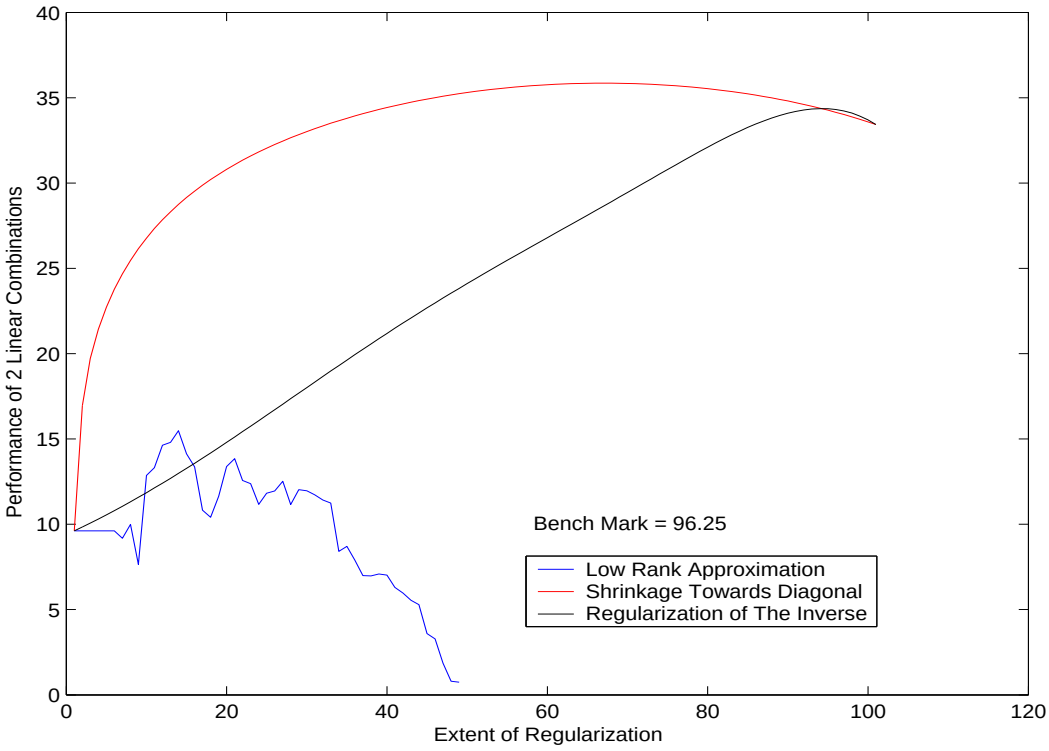


Figure 6.4.3: Performance of various regularization methods for Example 3

Example 4 In the fourth example, **W** and **B** were taken as in Example 3. However, instead of 10 observations, 20 observations were taken for each cluster. The performance of different regularization methods are displayed in the following figure.

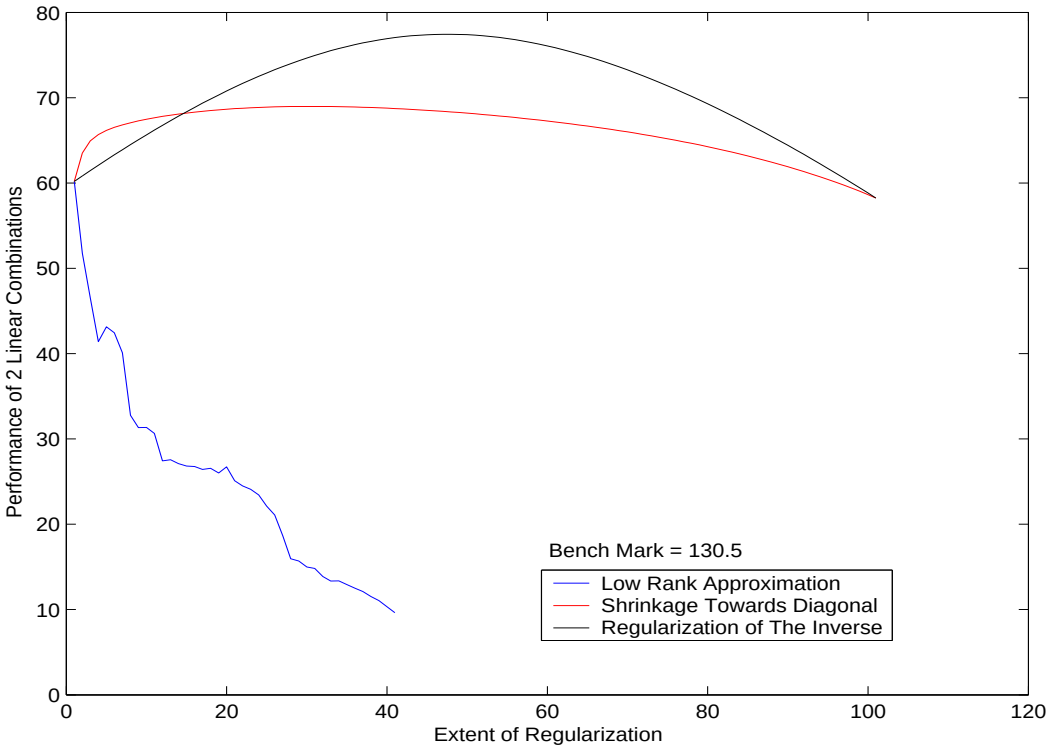


Figure 6.4.4: Performance of various regularization methods for Example 4

6.4.2 Discussion

We observe that when the within-cluster matrix $\mathbf{W}$ has very small eigenvalues, i.e., the condition number is very high, low rank approximation helps a lot to cope with the high dimensionality. In this situation, due to shortage of data, the eigenvectors of $\mathbf{W}$ corresponding to its small eigenvalues are very poorly estimated. Low rank approximation nullifies the contribution of such unstable estimates. So, in such a situation, low rank approximation is a good option. The maximizer of the trace of the ratio matrix, within the smaller subspace performs better. When the condition number is not very high, regularization of $\hat{\mathbf{W}}$ by shrinking it towards its diagonal version using a small shrinkage parameter, may improve the performance considerably. Shrinkage of $\hat{\mathbf{W}}$ towards its own diagonal version (again, with a small shrinkage parameter) also works well when the condition number is small.

6.5 An Example: Clustering of Tiger Pugmarks

6.5.1 Data Description

The problem is to estimate the number of tigers in a geographical area based on the pug-mark information collected during tiger census. Some training data are available. There are 37 features altogether. This is a high dimensional clustering problem as the dimension of the observations is comparable to the training sample size. In the training data, information about 33 tiger trails with one or more replications were available. Among them only 20 trails were known to correspond to distinct tigers. The total number of pugmarks from the 33 trails was 76.

6.5.2 Methodology Used

Initial analysis showed that 15 out of the 33 original variables either exhibit model heterogeneity or do not contain significant information to distinguish different tigers. So, we concentrated on the remaining 22 variables. The estimate $\hat{\mathbf{B}}$ was obtained by using the information on 20 different trails corresponding to distinct tigers, while $\hat{\mathbf{W}}$ was obtained by using information of all the 33 trails (some of which may correspond to the same tiger).

$\hat{\mathbf{W}}$ was observed to have a number of eigenvalues of the order 10^{-5}, whereas the maximum eigenvalue was .09. Moreover, the eigenvalues of the ratio matrix $\hat{\mathbf{R}} = \hat{\mathbf{W}}^{-1/2}\hat{\mathbf{B}}\hat{\mathbf{W}}^{-1/2}$ turned out to be widely spread. Eleven of them were less than 1/10 and three of them were of the order 10^{-11}. On the other hand, the maximum eigenvalue was 1256.2. Also, in the eigenvector of $\hat{\mathbf{R}}$ corresponding to its largest eigenvalue, as many as 8 elements are positive while the remaining 14 are negative. Since almost all the variables are physical dimensions, these are positively correlated. In such a situation, one would expect the elements of the said eigenvector to be mostly positive, as coefficients of the leading principal component must have the same sign whenever all elements of the dispersion matrix is positive.[1] All these observations about $\hat{\mathbf{W}}$ indicate a case of severe ill-conditioning. To cope with the problem of high dimensionality, low rank approximation of $\hat{\mathbf{W}}$ was considered. In this regularization, we replaced the 11 smallest eigenvalues by 0 and restricted our search for the optimal linear combinations within the eigenspace of $\hat{\mathbf{W}}$ corresponding to its 11 largest eigenvalues.

6.5.3 Results and Discussion

For given $d, 1 \leq d \leq 11$, the optimal d many linear combinations according to the $\Psi(\cdot)$ criteria is given by the d many orthogonal eigenvectors of $\hat{\mathbf{R}}_*$ corresponding to it s leading d many eigenvalues, where $\hat{\mathbf{R}}_* = \hat{\mathbf{W}}_*^{-1/2}\hat{\mathbf{B}}\hat{\mathbf{W}}_*^{-1/2}$ and $\hat{\mathbf{W}}_*^{-1/2}$ is the modified $\hat{\mathbf{W}}^{-1/2}$. In order to obtain the optimal number of linear combinations, we followed the conservative approach for average linkage. The probabilities $\Gamma_a(d, 1, \infty)$ for various d's are shown in the following table.

d	$\Gamma_a(d, 1, \infty)$
1	0.9714
2	0.9916
3	0.9958
4	0.9974
5	0.9981
6	0.9984
7	0.9986
8	0.9988
9	0.9987
10	0.9986
11	0.9985

Table 6.5.1: Performance of the optimal subsets of variables of different subset sizes

So, $d = 8$ maximizes $\Gamma_a(d, 1, \infty)$. Hence, the optimal set of linear combinations will be the first 8 orthogonal eigenvectors of $\hat{\mathbf{R}}_*$ corresponding to its leading 8 largest eigenvalues. The resulting clusters are displayed in Table 6.5.2. Here each "English letter" with several suffixes represent observations coming from the same cluster and different "English letters" correspond observations from different clusters.

$\{a_1\}$	$\{b_1\}$	$\{c_1\}$	$\{d_1\}$	$\{e_3\}$
$\{e_1, e_2, e_4, e_5, h_1\}$	$\{f_1, g_1, g_2\}$	$\{f_2, r_1\}$	$\{f_3, f_4\}$	$\{i_1\}$
$\{j_1\}$	$\{k_1\}$	$\{l_1\}$	$\{m_1\}$	$\{n_1\}$
$\{o_1, o_2\}$	$\{o_3, t_1, t_2, t_3, t_4, t_5\}$	$\{p_1, p_2, p_3, q_1\}$	$\{q_2\}$	$\{s_1\}$

Table 6.5.2: Clusters of the training data using 8 optimal linear combinations

Bibliography

[1] Bellman, R. (1960). *Introduction to matrix analysis*, McGraw-Hill, New York-Toronto-London.

[2] Bensmail, H. and Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition, *Journal of the American Statistical Association* **91**, 1743–1748.

[3] Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**, 123–140.

[4] Bickel, P.J. and Levina, E.V. (2004). Some theory for Fisher's linear discriminant fuction, 'naive Bayes', and some alternetives when there are many more variables than observations, *Bernoulli* **10**, 989–1010.

[5] Dey, D.K. and Srinivasan, C., (1985). Estimation of a covariance matrix under Stein's loss, *The Annals of Statistics* **137**, 1581–1591.

[6] Effron, B. and Morris, C. (1976). Multivariate empirical Bayes and Estimation of covariance matrices, *The Annals of Statistics* **4**, 22–32.

[7] Fowlkes, E.B., Gnanadesikan, R., and Kettering, J.R. (1988). Variable selection in clustering, *Journal of Classification* **5**, 205–228.

[8] J. Friedman (1989). Regularised discriminant analysis, *Journal of the American Statistical Association* **84**, 165–175.

[9] Friedman, J.H., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion), *The Annals of Statistics* **28**, 337–407.

[10] Ghosh, A.K., Chaudhuri, P. and Sengupta, D. (2006). Classification using kernel density estimates: multi-scaling analysis and visualization, *Technometrics* **48**, 120–132.

[11] Ghosh, D., and Chinnaiyan, A.M. (2002). Mixture modelling of gene expression data from microarray experiments, *Bioinformatics* **18**, 275–286.

[12] Haff, L.R. (1980). Empirical Bayes estimation of the multivariate normal covariance matrix, *The Annals of Statistics* **8**, 586–597.

[13] James, M. and Stein C. (1961). Estimation with quadratic loss. In *Proceedings of The Fourth Berkeley Symposium on Mathematical, Statistical and Probability* **Vol. 1**, 361–379.

[14] Lin, S.P. and Perlman, M.D. (1985). A Monte Carlo comparison of four estimators of a Covariancd Matrix. In *Multivariate Analysis* **IV**, P.R. Krishnaiah, ed., North Holland, Amsterdam, 411–429.

[15] McLachlan, G.J., Bean, R.W. and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data, *Bioinformatics* **18**, (413–422)

[16] Schaafsma, W. (1982). Selecting variables in discriminant analysis for improving upon classical procedures. In *Handbook of Statistics,* ***Vol. 2*** *(Classification, Pattern Recognition and Reduction of Dimensionality)*, P.R. Krishnaiah and L.N. Kanal, eds., North-Holland, Amsterdam, 857–881.

[17] Schapire, R.E., Fruend, Y., Bartlett, P. and Lee, W. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods, *The Annals of Statistics* **26**, 1651–1686

[18] Tadesse, M.G., Sha, N. and Vannucci, M. (2005). Bayesian variable selection in clustering high dimensional data, *Journal of the American Statistical Association* **100**, 602–617.