2005

Decision Theory Classification Of High-dimensional Vectors Based On Small Samples

David Bradshaw
University of Central Florida

DECISION THEORY CLASSIFICATION OF HIGH-DIMENSIONAL VECTORS BASED ON SMALL SAMPLES

by

DAVID J. BRADSHAW

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in the Department of Mathematics
in the College of Arts and Sciences at the University of Central Florida
Orlando, Florida

Fall Term
2005

Major Professor: Dr. Marianna Pensky

ABSTRACT

In this paper, we review existing classification techniques and suggest an entirely new procedure for the classification of high-dimensional vectors on the basis of a few training samples. The proposed method is based on the Bayesian paradigm and provides posterior probabilities that a new vector belongs to each of the classes, therefore it adapts naturally to any number of classes. Our classification technique is based on a small vector which is related to the projection of the observation onto the space spanned by the training samples. This is achieved by employing matrix-variate distributions in classification, which is an entirely new idea. In addition, our method mimics time-tested classification techniques based on the assumption of normally distributed samples. By assuming that the samples have a matrix-variate normal distribution, we are able to replace classification on the basis of a large covariance matrix with classification on the basis of a smaller matrix that describes the relationship of sample vectors to each other.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION AND FORMULATION OF THE PROBLEM

1.1 Introduction

The problem of pattern classification has been around for almost a century, and with recent technological developments, it is experiencing a renewed enthusiasm. The problem itself is quite simple to describe: "What's the difference between *these* and *those*?" Or, more to the point, "is *that* more like one of *these* or one of *those*?" The problem is to construct an algorithm, called a *decision rule*, which classifies future observations as coming from C predetermined classes. In order to devise the decision rule, one is given a set of N observations, called *training samples*, which come from known classes. Using the training samples, the classifier "learns" a distinctive description of each class. Depending on the design of the classifier, the learning process can involve estimating probability distributions (decision theoretic approach), finding coefficients of a separating hyperplane (linear discriminants), or various other techniques. The rule is usually designed on the basis of some optimization criterion, for example, minimizing the percentage of future misclassifications.

We can trace the origins of pattern classification to a 1936 study by Ronald Fisher of

the differences between various types of Iris plants, *Iris versicolor* and *Iris setosa*. In the experiment, Fisher represented each plant by its *feature vector* $\mathbf{x} = (x_1, ..., x_4)^T$, where $x_1 =$ sepal length, $x_2 =$ sepal width, $x_3 =$ petal length, and $x_4 =$ petal width, all in centimeters. He sought a coefficient vector $\mathbf{w} = (w_1, ..., w_4)^T$ such that the values of the linear function $\mathbf{w}^T\mathbf{x}$ for each species would be widely separated. Fisher argued that the coefficient vector $\mathbf{w}$ should be chosen to maximize the ratio of the square of the mean difference to the within-group variance (see Section 2.3.2). If we denote the training samples from group i by $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,N_i},\ i = 1, 2$, and the sample means by $\mathbf{m}_i$, then the coefficients can be found by maximizing

$$J = \frac{(\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2)^2}{\mathbf{w}^T\mathbf{S}_W\mathbf{w}}, \tag{1.1}$$

where $\mathbf{S}_W = \sum_{i=1}^{2}\sum_{j=1}^{N_i}(\mathbf{x}_{i,j} - \mathbf{m}_i)(\mathbf{x}_{i,j} - \mathbf{m}_i)^T$ is proportional to the within-group pooled covariance matrix. The coefficients are then given by

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2). \tag{1.2}$$

(Note that the parameters w_i of the classifier were "learned" from the training samples. Classifiers that work in this way are often called *learning machines*, and setting the parameters based on the samples is called *training*.) A natural decision rule here would be to pick some suitable constant k (between $\mathbf{w}^T\mathbf{m}_1$ and $\mathbf{w}^T\mathbf{m}_2$) and classify a new vector $\mathbf{x}$ according to whether $\mathbf{w}^T\mathbf{x}$ is greater than or less than k. This decision rule is known as *Fisher's Linear Discriminant Function*.

Fisher's study was a two-class discrimination, where the dimension of the vector space n was small compared to the number of available training samples N. In this paper, we are interested in the opposite situation. That is, given C classes and N training samples in $\mathbb{R}^n$

with $N << n$, determine an appropriate classification rule. This situation is becoming more common as our technological capacity grows. For example, even a low resolution digital photograph will have thousands of pixels, and most image classification applications have a limited number of training images. In the medical field, several studies have demonstrated great potential in classifying various types of tumors based on microarray measurements for thousands of genes. These *HDLSS* (high dimension, low sample-size) problems do not cooperate with the traditional classification methods.

The statistical methods of classification all rely on some sort of density estimation. Non-parametric estimation will not work in HDLSS problems, because the number of training samples needed grows exponentially with dimension. When a parametric form is assumed (generally normal), the number of parameters to be estimated is $O(n^2)$. Linear discriminant functions are less computationally complex, and hence do not suffer as much from the "curse of dimensionality," however they do not support the Bayesian paradigm (except in special cases). In this paper, we propose a classification method designed for the HDLSS problem that combines the Bayesian decision rules with linear discriminant functions

In Fisher's study, a feature vector consisted of petal and sepal dimensions, but there were other features that could have improved results, like height or mass or leaf-color. Our intuition tells us that the more characteristics we use to distinguish, the more accurate our classification will be. But this is not always the case. For example, adding height and mass to the feature vector may have improved performance, but because of the strong statistical dependence of the two, the additional information might have some redundancy, and the performance increase could have been achieved with only one or the other. Hence care should be given to feature selection, not only to keep the feature vector to a reasonable size, but also

to ensure that the chosen features separate the categories well. In classification applications, the features are usually selected by a scientist or someone with intimate knowledge of the area. This was the case in Fisher's experiment, because he was an evolutionary biologist and statistician, so he was familiar with the distinguishing characteristics of various Iris plants.

While we will be considering only a subset of the general theory of pattern classification, we can mention some of the areas that fall outside of our scope. Our discussion will only deal with *supervised learning* (where the category of each training sample is known), but there are many reasons to study *unsupervised learning*. As an example, consider training a speech-recognition classifier, where samples of recorded speech are free, but labeling each word would be a time-consuming process. In this case, one might desire a classifier that can learn from large amounts of unlabeled data, and perhaps be fine-tuned with a smaller amount of labeled data. Or in data mining applications, the contents of a large database might be unknown, and it is desired to analyze the data for groups of patterns whose members are similar to each other but distinct from other groups. This type of classification, sometimes called *data clustering*, has a wide range of applications, as does the area of *reinforced learning*, where the machine is trained by labelling its decisions as right or wrong.

As in Fisher's experiment, this paper will deal with classification of vectors in Euclidian space. But there are also various non-metric methods, where the *feature space* (the space containing the set of all feature vectors, $\mathbb{R}^4$ in Fisher's experiment) lacks a natural notion of distance, similarity, or ordering of its elements. This is usually the case when the problem involves nominal data, such as word-based descriptions. In cases like these we can construct *decision trees* and use a "twenty-questions" like approach to classification. There are even rule-based and grammatical classifiers that can be used when there is an underlying structure

to the elements.

In what follows we shall denote vectors by boldface lowercase letters and matrices by boldface capital letters. We shall keep the notations commonly used in matrix algebra, i.e. $|A|$ is the determinant of the matrix A, $\|\mathbf{x}\|$ is the Euclidean norm of the vector $\mathbf{x}$, etc. Let $\omega_1, \ldots, \omega_C$ represent the C states of nature, or classes. We will use $\mathbf{x}_{i,j} \in \mathbb{R}^n$ to denote the j^{th} training sample from class ω_i, $j = 1, \ldots, N_i$, where the total number of training samples is $N = \sum_{i=1}^{C} N_i$. The objective is to assign a new vector $\mathbf{x} \in \mathbb{R}^n$ to one of the classes ω_i based on the training samples $\{\mathbf{x}_{i,j}\}$. We will summarize some of the current methods of classification, including generalizations and shortcomings. We will then propose a new method designed for classification of large-dimensional vectors on the basis of small number of samples ($n >> N$).

2

BACKGROUND INFORMATION

2.1 Bayesian Classification Methods

2.1.1 Bayesian decision rules

Let us first introduce the Bayesian approach to classification. Since the classifier should employ all of the prior knowledge of the problem, it naturally depends on what is known beforehand. Consider the simplest case where we know only the *a priori* class probabilities $p(\omega_i)$. If our goal is to minimize the probability of error, then we would choose the class with the highest value for $p(\omega_i)$. While this would achieve the goal (based on the limited information), the classification decision is independent of the measurement vector $\mathbf{x}$. To improve our chances, we should compute the *posterior* probability that the given vector is in class ω_i. If we know the class-conditional probabilities $p(\mathbf{x}|\omega_i)$ (or perhaps estimated them from the training samples), then in order to minimize the error, we need to choose the class with the highest posterior probability $p(\omega_i|\mathbf{x})$. From Bayes theorem, we have

$$\begin{aligned} p(\omega_i|\mathbf{x}) &= \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{\sum_{j=1}^{C} p(\mathbf{x}|\omega_j)p(\omega_j)}. \end{aligned}$$

Observe that the denominator is merely a scaling factor and is independent of i. Thus, Bayes decision rule for minimum error is

$$p(\mathbf{x}|\omega_j)p(\omega_j) \geq p(\mathbf{x}|\omega_k)p(\omega_k),\ \forall k \implies \mathbf{x} \in \omega_j. \tag{1.1}$$

In the event of equal posterior probabilities, the decision is made based on the class priors.

To see that the Bayes decision rule minimizes the probability of error, let Ω_i denote the set of points in $\mathbb{R}^n$ which are assigned to class ω_i, $i = 1, \ldots, C$. Then the probability of error can be written

$$\begin{aligned} p(\text{error}) &= \sum_{i=1}^{C} p(\omega_i)p(\text{error}|\omega_i) \\ &= \sum_{i=1}^{C} p(\omega_i) \int_{\mathbb{R}^n \setminus \Omega_i} p(\mathbf{x}|\omega_i)\ \mathrm{d}\mathbf{x} \\ &= \sum_{i=1}^{C} p(\omega_i) \left(1 - \int_{\Omega_i} p(\mathbf{x}|\omega_i)\ \mathrm{d}\mathbf{x}\right) \\ &= 1 - \sum_{i=1}^{C} \int_{\Omega_i} p(\mathbf{x}|\omega_i)p(\omega_i)\ \mathrm{d}\mathbf{x}. \end{aligned}$$

Therefore, minimizing the probability of error is achieved by choosing Ω_i to be precisely the region where $p(\mathbf{x}|\omega_i)p(\omega_i)$ is the largest out of all the classes. This leads to a convenient representation for the minimum possible error rate, or Bayes error rate,

$$e_B = 1 - \sum_{i=1}^{C} \int_{\mathbb{R}^n} \max_i \left[p(\mathbf{x}|\omega_i)p(\omega_i)\right]\ \mathrm{d}\mathbf{x}.$$

2.1.2 Minimum Risk Criterion

Often the minimum error rate criterion is not appropriate in practice because certain misclassifications are more costly than others. For example, in a medical diagnosis problem it is more dangerous to misclassify a sick patient as healthy than vice versa. In this situation,

rather than designing a classifier to achieve the minimum error rate, we should assign a cost to each misclassification and design a classifier to minimize the expected cost, or *risk*. To this end, we denote by $\lambda_{j,i}$ the cost of misclassifying a pattern from class ω_j into ω_i. The *conditional risk* of assigning a pattern $\mathbf{x}$ to ω_i is then given by

$$l_i(\mathbf{x}) = \sum_{j=1}^{C} \lambda_{j,i} p(\omega_j|\mathbf{x}).$$

As before, we let Ω_i be the region in $\mathbb{R}^n$ that is classified as ω_i. Then the average risk over region Ω_i is

$$\begin{aligned} r_i &= \int_{\Omega_i} l_i(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \\ &= \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{j,i} p(\omega_j|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Summing over all the regions, we get the overall risk,

$$r = \sum_{i=1}^{C} \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{j,i} p(\omega_j|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{1.2}$$

The above expression for risk will be minimized if Ω_i is exactly the set of points $\mathbf{x}$ where the integrand is minimized. Thus, Bayes decision rule for minimum risk is

$$\sum_{j=1}^{C} \lambda_{j,i} p(\omega_j|\mathbf{x}) \leq \sum_{j=1}^{C} \lambda_{j,k} p(\omega_j|\mathbf{x}), \;\; k = 1, \ldots, C \implies \mathbf{x} \in \omega_i, \tag{1.3}$$

and the minimum possible risk is

$$r^* = \int_{\mathbb{R}^n} \min_i \left[\sum_{j=1}^{C} \lambda_{j,i} p(\omega_j|\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x} \right].$$

As a special case, if we assign unit cost to misclassifications, and zero cost to correct classification, i.e. $\lambda_{j,i}$ is the Kronecker delta $\delta_{j,i}$, then the decision rule becomes the Bayes rule for minimum error.

2.1.3 Minimax criterion

The Bayes classification rules depend on the prior class probabilities $p(\omega_i)$ and the within-class distributions $p(\mathbf{x}|\omega_i)$. But sometimes the situation calls for a decision rule that will work well for a range of prior class probabilities, such as when the relative frequency of new objects to be classified varies throughout the year. In this case, the Bayes minimum risk classifier for one value of the priors might lead to an unacceptably high risk as the priors fluctuate (assuming the decision regions Ω_i remain fixed). In this case a *minimax* criterion can be used to minimize the maximum possible risk over the range of prior class probabilities. To illustrate, consider a two-class problem. From (1.2), and noting that $p(\omega_j|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\omega_j)p(\omega_j)$, the Bayes risk is

$$\begin{aligned} r &= \sum_{i=1}^{2} \int_{\Omega_i} \sum_{j=1}^{2} \lambda_{j,i} p(\mathbf{x}|\omega_j) p(\omega_j) \, \mathrm{d}\mathbf{x} \\ &= \int_{\Omega_1} \left[\lambda_{1,1} p(\mathbf{x}|\omega_1) p(\omega_1) + \lambda_{2,1} p(\mathbf{x}|\omega_2) p(\omega_2) \right] \, \mathrm{d}\mathbf{x} \\ &\quad + \int_{\Omega_2} \left[\lambda_{1,2} p(\mathbf{x}|\omega_1) p(\omega_1) + \lambda_{2,2} p(\mathbf{x}|\omega_2) p(\omega_2) \right] \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Using the fact that $p(\omega_2) = 1 - p(\omega_1)$, we can write r as a function of $p(\omega_1)$ alone

$$\begin{aligned} r = &\int_{\Omega_1} \lambda_{2,1} p(\mathbf{x}|\omega_2) \, \mathrm{d}\mathbf{x} + \int_{\Omega_2} \lambda_{2,2} p(\mathbf{x}|\omega_2) \, \mathrm{d}\mathbf{x} \\ &+ p(\omega_1) \left[\left(\int_{\Omega_1} \lambda_{1,1} p(\mathbf{x}|\omega_1) - \lambda_{2,1} p(\mathbf{x}|\omega_2) \, \mathrm{d}\mathbf{x} \right) + \left(\int_{\Omega_2} \lambda_{1,2} p(\mathbf{x}|\omega_1) - \lambda_{2,2} p(\mathbf{x}|\omega_2) \, \mathrm{d}\mathbf{x} \right) \right]. \end{aligned}$$

Since this is a two-category case, we can use $\int_{\Omega_2} + \int_{\Omega_1} = 1$, obtaining

$$\begin{aligned} r = &\lambda_{2,2} + (\lambda_{2,1} - \lambda_{2,2}) \int_{\Omega_1} p(\mathbf{x}|\omega_2) \, \mathrm{d}\mathbf{x} \\ &+ p(\omega_1) \left[(\lambda_{1,1} - \lambda_{2,2}) + (\lambda_{1,2} - \lambda_{1,1}) \int_{\Omega_2} p(\mathbf{x}|\omega_1) \, \mathrm{d}\mathbf{x} - (\lambda_{2,1} - \lambda_{2,2}) \int_{\Omega_1} p(\mathbf{x}|\omega_2) \, \mathrm{d}\mathbf{x} \right]. \end{aligned}$$

The latter shows that, for fixed decision regions Ω_i, risk is linear in $p(\omega_1)$. Note that Bayes risk is not linear in $p(\omega_1)$, because the Bayes decision regions would not remain fixed. Hence

fixing the decision regions according to (1.3) for some value of $p(\omega_1)$, then varying $p(\omega_1)$ results in a linear change in risk. If we can find decision regions such that this constant of proportionality is zero, i.e.

$$(\lambda_{1,2} - \lambda_{1,1}) \int_{\Omega_2} p(\mathbf{x}|\omega_1)\ \mathrm{d}\mathbf{x} - \lambda_{2,2} = (\lambda_{2,1} - \lambda_{2,2}) \int_{\Omega_1} p(\mathbf{x}|\omega_2)\ \mathrm{d}\mathbf{x} - \lambda_{1,1},$$

then we will achieve the minimax risk

$$r_{mm} = \lambda_{2,2} + (\lambda_{2,1} - \lambda_{2,2}) \int_{\Omega_1} p(\mathbf{x}|\omega_2)\ \mathrm{d}\mathbf{x},$$

which is the maximum value of the Bayes risk for $p(\omega_1) \in [0, 1]$.

2.2 Linear Discriminating Functions

The discrimination methods described above (except Fisher's Linear Discriminant) are all based on the underlying class-conditional probability densities, and the training samples are used to estimate these pdf's. The decision rule was then determined by the appropriate method (minimizing error, minimax criterion, etc). We will now introduce a different approach, called a *linear discriminant function*, which bypasses calculation of the pdf's and attempts to find an appropriate decision rule directly from the training samples. The fundamental assumption for these classifiers is that the decision rule can be expressed in terms of discriminant functions which are linear in the components of $\mathbf{x}$ (or linear in some set of functions of components of $\mathbf{x}$). The problem is then to choose the coefficients of the functions to provide a certain optimality in the training samples. As in the decision theoretic approach, this is generally done by optimizing some criterion function. One problem, however, is that it is difficult to derive minimum-risk or minimum-error criterion functions, so we generally use functions that are easier to deal with, such as J_p or J_q (see below).

The main benefit of the linear discriminant function over the decision theoretic approach is its simplicity. First, the learning process involves specifying far fewer parameters. A linear function of a feature vector $\mathbf{x} \in \mathbb{R}^n$ has only $(n+1)$ parameters, while estimating the mean vector and covariance matrix of a Gaussian involves $n(n+3)/2$ parameters. Second, many efficient computer algorithms have been developed to minimize various criterion functions. Hence, for applications with high dimensional feature spaces, such as image classification and certain medical problems, linear discriminant functions are a good choice.

To illustrate this method, consider $C = 2$ classes. In the linear discriminant approach, we seek a function of the form

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0,$$

where $\mathbf{w}$ is called the *weight vector*, and w_0 the *threshold weight*, such that $g(\mathbf{x}) > 0$ for ω_1 and $g(\mathbf{x}) < 0$ for ω_2. Geometrically, this amounts to splitting the feature space with a hyperplane $\mathbf{w}^T\mathbf{x} + w_0 = 0$, where $\mathbf{w}$ is the normal vector, and $w_0/\|\mathbf{w}\|$ is the distance from the origin. Classifying a new vector is equivalent to deciding on which side of the hyperplane the vector lies.

While the decision theoretic approach generalizes well to the C-class problem, the linear discriminant function is inherently a binary classifier, because a hyperplane can only split the feature space into two regions. However, there are various methods to generalize any binary classifier to C classes. One method is to construct $C(C-1)/2$ linear discriminants, one for each pair of classes. Another is to reduce the problem to C two-class problems, where the i^{th} problem discriminates between ω_i and not ω_i. Unfortunately, both of these methods can lead to ambiguously defined regions. What is often done for linear discriminants is a

variation of the latter: construct C linear discriminant functions

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0},\ i = 1, \ldots, C, \tag{2.4}$$

and assign $\mathbf{x}$ to the class with the largest value of $g_i(\mathbf{x})$ (ignoring ties). The resulting classifier divides the feature space into C regions, where the border between two neighboring regions Ω_i and Ω_j is a portion of the hyperplane

$$(\mathbf{w}_i^T - \mathbf{w}_j^T)\mathbf{x} = w_{j0} - w_{i0}.$$

To illustrate the connection between the decision theoretic approach and the linear discriminant function, consider the two-class problem where the class-conditional densities are multivariate normal with different means but equal covariance matrices, i.e. $P(\mathbf{x}|\omega_1) \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $P(\mathbf{x}|\omega_2) \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, so that the pdf's are of the form

$$P(\mathbf{x}|\omega_i) = (2\pi)^{-n/2} \left|\boldsymbol{\Sigma}\right|^{-1/2}\ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)},\ i = 1, 2.$$

This model often appears in applications (such as signal detection) where the observed feature vector $\mathbf{x}$ represents a class-prototype vector ($\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$) corrupted by noise (atmospheric interference, instrumental error, etc.). If we use the Bayes decision rule for minimum error given in (1.1), then we assign $\mathbf{x}$ to class ω_1 if and only if

$$\frac{p(\omega_1)}{(2\pi)^{n/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)} > \frac{p(\omega_2)}{(2\pi)^{n/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\ e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)}.$$

Simplifying the last inequality and taking logarithms of both sides, we arrive at

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) > \ln \frac{p(\omega_2)}{p(\omega_1)} - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2).$$

Cancelling the quadratic term $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ in both sides of the inequality and further simplifying, we obtain a discrimination rule that is linear in $\mathbf{x}$:

$$\mathbf{x} \in \omega_1 \iff \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) > 0,$$

where

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$
$$\mathbf{x}_0 = \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} - \frac{\ln[p(\omega_1)/p(\omega_2)]}{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

Thus, under these fairly general assumptions, the Bayes minimum error rate can only be achieved by a linear discriminant function. Note that when the priors are equal, the decision surface will pass through the midpoint of the means. For unequal priors, the surface shifts away from the more likely mean. For more extreme values of the priors, it may not even pass between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

In this example, and in fact in any C-class linear discriminant function, the boundary between two adjacent regions will be a portion of a hyperplane. In particular, each region will be connected. This tends to make linear machines especially suited for problems where the class-conditional densities are unimodal. However, for some multimodal densities, or even Gaussians with equal means but different covariance matrices, the optimal Bayes regions are not connected, so linear machines do not perform well in these circumstances.

2.2.1 Generalized Linear Discriminant Functions

When classes cannot be separated efficiently by hyperplanes, the linear machines can be modified by adding additional terms involving products of pairs of components of $\mathbf{x}$, so that instead of (2.4), the discriminant functions take the form

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} x_i x_j,$$

called a *quadratic discriminant function*. This gives the added flexibility of using various quadratic surfaces to separate regions, but at a computational expense when n is large.

Higher degree terms can be added, which can be though of as truncated series expansions of some arbitrary (nonlinear) $g(\mathbf{x})$. While these function are not necessarily linear in components of $\mathbf{x}$, they are linear in the components of some vector-valued function $\mathbf{y}(\mathbf{x}) \in \mathbb{R}^{\hat{n}}$. For example, when $n = 2$, the quadratic discriminant function above can be thought of as a linear function of $\mathbf{y} = (1, x_1, x_2, x_1x_2) \in \mathbb{R}^4$. For polynomial discriminants, the $\hat{n}$ functions are all the possible monomials of the components of $\mathbf{x}$, but for more general cases they may be computed by some feature detecting system. The mapping from $\mathbf{x}$ to $\mathbf{y}$ merely reduces the problem to finding a linear discriminant function. For true linear discriminant functions, a common practice is to set $\mathbf{y}^T = (1, x_1, \ldots, x_n)$ so that (2.4) can be written as $g_i(\mathbf{x}) = \mathbf{w}^T\mathbf{y}$ for some $\mathbf{w} \in \mathbb{R}^{n+1}$. The vectors $\mathbf{y}$ and $\mathbf{w}$ are often called the *augmented feature vector* and *augmented weight vector*, respectively.

2.2.2 Criterion Functions

We now consider various methods to determine the parameters in linear discriminant functions. Without loss of generality, we will consider the two-class problem where $g(\mathbf{y}) = \mathbf{w}^T\mathbf{y}$, and $\mathbf{y}$ is the image of $\mathbf{x}$ in $\mathbb{R}^{\hat{n}}$. If the training samples $\{\mathbf{y}_{i,j},\ j = 1, \ldots, N_i, i = 1, 2\}$ are *linearly separable*, i.e. they can be separated by a hyperplane without error, then the *solution vector* $\mathbf{w}$ (which is usually not unique) should satisfy the inequalities

$$\mathbf{w}^T\mathbf{y}_{1,j} > 0,\ j = 1, \ldots, N_1,$$

$$\mathbf{w}^T\mathbf{y}_{2,j} < 0,\ j = 1, \ldots, N_2.$$

But rather than dealing with mixed inequalities, we can make the trivial transformation $\mathbf{y}_{2,j} \to -\mathbf{y}_{2,j},\ j = 1, \ldots, N_2$, so that an error-free separation occurs if and only if

$$\mathbf{w}^T\mathbf{y}_{i,j} > 0,\ j = 1, \ldots, N_i,\ i = 1, 2. \tag{2.5}$$

Often, the training samples are not linearly separable, so a solution to (2.5) might not be possible. In these cases, we can specify a criterion function $J(\mathbf{w}; \mathbf{y}_{1,1}, \ldots, \mathbf{y}_{1,N_1}, \mathbf{y}_{2,1}, \ldots, \mathbf{y}_{2,N_2})$ to be minimized (or maximized) with respect to $\mathbf{w}$. The criterion function $J(\mathbf{w})$ is usually difficult to optimize analytically, so this process is done by computer calculations.

A natural choice for the criterion function $J(\mathbf{w})$ is to set it equal to the number of samples misclassified by $\mathbf{w}$. However, this function is piecewise constant, so it is not a good choice for optimization algorithms. A better function to work with is the *Perceptron criterion function*

$$J_p(\mathbf{w}) = \sum_{\mathbf{y} \in \mathcal{Y}} (-\mathbf{w}^T \mathbf{y}),$$

where $\mathcal{Y}$ is the set of samples misclassified by $\mathbf{w}$. Since $\mathbf{w}^T \mathbf{y} < 0$ whenever $\mathbf{y}$ is misclassified, then J_p is nonnegative. Geometrically, J_p is proportional to the sum of the distances from the separating hyperplane to the misclassified training samples samples, so it is zero only when it separates without error. This criterion function lends itself well to a gradient decent procedure, because $\nabla J_p = \sum_{\mathbf{y} \in \mathcal{Y}} (-\mathbf{y})$. Thus, a gradient decent algorithm for minimizing the Perceptron criterion function can be summarized as follows: the new weight vector equals the old weight vector plus some multiple (increment size) of the sum of the misclassified samples. It is shown in Duda et al. [2001] that in the linearly separable case, a fixed-increment gradient decent procedure will always terminate at a separating hyperplane for any initial choice of $\mathbf{w}$. However, in the linearly separable case, there is an entire wedge-shaped *solution region* which gives error-free separation. To ensure that the solution vector lies near the middle of this region (which we might expect would lead to better performance), we can introduce a *margin* $b > 0$, so that $\mathbf{y} \in \mathcal{Y}$ whenever $\mathbf{w}^T \mathbf{y} \leq b$. In that case, a vector $\mathbf{y}$ would be "misclassified" whenever $\mathbf{w}^T \mathbf{y}$ does not exceed the margin.

An alternative to the Perceptron is the related function

$$J_q(\mathbf{w}) = \sum_{\mathbf{y}\in\mathcal{Y}} (\mathbf{w}^T\mathbf{y})^2.$$

The main benefit is that J_q is smoother and has a continuous gradient. However, a serious disadvantage is that J_q is so smooth near the boundary of the solution region that the sequence of weight vectors can converge to a point on the boundary, e.g. zero vector. Furthermore, J_p and J_q can both be overly dominated by the largest sample vectors. The latter can be corrected by using the criterion function

$$J_r(\mathbf{w}) = \frac{1}{2}\sum_{\mathbf{y}\in\mathcal{Y}} \frac{(\mathbf{w}^T\mathbf{y} - b)^2}{\|\mathbf{y}\|^2},$$

where $\mathcal{Y}$ is the set of samples for which $\mathbf{w}^T\mathbf{y} \leq b$.

2.2.3 Minimum Squared Error Criterion

The above methods deal with criterion functions that only use the misclassified samples. We now consider a technique that uses all of the training samples. The main idea is to replace the inequality constraints $\mathbf{w}^T\mathbf{y}_{i,j} > 0$ with the equalities $\mathbf{w}^T\mathbf{y}_{i,j} = b_{i,j}$. Then the problem of solving a set of inequalities becomes a more straightforward problem of solving a system of linear equations.

In matrix form, the objective is to find a weight vector $\mathbf{w} \in \mathbb{R}^{\hat{n}}$ such that

$$\mathbf{Y}\mathbf{w} = \mathbf{b},$$

where $\mathbf{Y}$ is the $N \times \hat{n}$ matrix whose rows are the training samples, and $\mathbf{b}$ is a vector which we specify. Matrix $\mathbf{Y}$ is generally not square, so an exact solution is usually not possible. Suppose $N > \hat{n}$, so that $\mathbf{w}$ is over-determined. In this case, we can find the least squares

solution for $\mathbf{w}$, i.e. the one which minimizes

$$J_s(\mathbf{w}) = \|\mathbf{Y}\mathbf{w} - \mathbf{b}\|^2.$$

A necessary condition for minimization is obtained by setting the gradient $\nabla J_s = 2\mathbf{Y}^T(\mathbf{Y}\mathbf{w} - b)$ equal to zero, so that

$$\mathbf{Y}^T\mathbf{Y}\mathbf{w} = \mathbf{Y}^T\mathbf{b}. \tag{2.6}$$

If $\mathbf{Y}^T\mathbf{Y}$ is invertible, then this gives us the Minimum Squared-Error (MSE) Solution because it minimizes the norm of the error vector $\mathbf{e} = \mathbf{Y}\mathbf{w} - \mathbf{b}$.

To see a connection between the MSE solution and Fisher's Linear Discriminant, suppose we partition $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{N_1} & \mathbf{X}_1 \\ \mathbf{1}_{N_2} & \mathbf{X}_2 \end{bmatrix},$$

where $\mathbf{X}_i$ is the $N_i \times \hat{n}$ matrix whose rows are the training samples from class ω_i, and $\mathbf{1}_{N_i} \in \mathbb{R}^{N_i}$ is a column vector of 1's. We partition $\mathbf{w}$ and $\mathbf{b}$ accordingly, as

$$\mathbf{w} = \begin{bmatrix} w_0 \\ \mathbf{w}_v \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} N/N_1\ \mathbf{1}_{N_1} \\ -N/N_2\ \mathbf{1}_{N_2} \end{bmatrix}.$$

Thus we seek the least squares solution to the system of equations

$$\mathbf{w}^T\mathbf{y}_{1,j} = N/N_1, \quad j = 1, \ldots, N_1,$$

$$\mathbf{w}^T\mathbf{y}_{2,j} = -N/N_2,\ j = 1, \ldots, N_2,$$

where $\mathbf{y}_{i,j}$ is j^{th} the augmented feature vector from class ω_i. Plugging these block forms into equation (2.6) and simplifying yields

$$\begin{bmatrix} N & (N_1\mathbf{m}_1 + N_2\mathbf{m}_2)^T \\ N_1\mathbf{m}_1 + N_2\mathbf{m}_2 & \mathbf{S}_W + N_1\mathbf{m}_1\mathbf{m}_1^T + N_2\mathbf{m}_2\mathbf{m}_2^T \end{bmatrix} \begin{bmatrix} w_0 \\ \mathbf{w}_v \end{bmatrix} = \begin{bmatrix} 0 \\ N(\mathbf{m}_1 - \mathbf{m}_2) \end{bmatrix}.$$

Here

$$\mathbf{m}_i = \sum_{j=1}^{N_i} \mathbf{y}_{i,j},\ i = 1, 2,$$
$$\mathbf{S}_W = \sum_{i=1}^{2} \sum_{j=1}^{N_i} (\mathbf{y}_{i,j} - \mathbf{m}_i)(\mathbf{y}_{i,j} - \mathbf{m}_i)^T,$$

are the group means and within-class *scatter* matrix (which is just $N - 2$ times the sample covariance matrix), respectively. The top equation leads to $w_0 = -\mathbf{m}^T \mathbf{w}_v$, where $\mathbf{m}$ is the mean of all the samples. Plugging this in to the bottom equation and simplifying, we get

$$\left[\mathbf{S}_W + \frac{N_1 N_2}{N}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T\right] \mathbf{w}_v = N(\mathbf{m}_1 - \mathbf{m}_2).$$

Note that $\frac{N_1 N_2}{N}(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w}$ can be written as $\alpha(\mathbf{m}_1 - \mathbf{m}_2)$, where α is a scalar that depends on $\mathbf{w}$. Simplifying and solving, we get $\mathbf{w}_v = \mathbf{S}_W^{-1}(N - \alpha)(\mathbf{m}_1 - \mathbf{m}_2)$. But since only the direction of $\mathbf{w}_v$ is important, the separating hyperplane in $\mathbb{R}^n$ is determined by its normal vector $\mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$, which is the same as in Fisher's Linear Discriminant, equation (1.2).

2.2.4 Support Vector Machines

In the last several decades, a new type of linear classifier has been developed, called the Support Vector Machine (SVM). It is similar to the ones mentioned above in that it searches for an optimal hyperplane with respect to some criterion function. However, in this case is it difficult to write explicitly. Furthermore, the separable and non-separable cases have slightly different derivations. The basic idea is to map the vectors $\mathbf{x}_{i,j}$ into a high-dimensional vector space (which is done implicity, so the actual mapping need not be computed) and search for a separating hyperplane which gives the largest margin between the two classes. The concept of a margin was introduced with the Perceptron criterion function, where it is assumed that

a wider margin will result in better performance of the classifier. Like all linear classifiers, SVMs are designed as binary classifiers. However, this relatively new area has been a popular topic for research, including a formulation for the multicategory case [Lee et al., 2004].

We will start with the simple case where the training samples are linearly separable. For notational convenience, let us temporarily drop the double subscripts and refer the sample vectors simply as $\mathbf{x}_1, \ldots, \mathbf{x}_N$, where $\mathbf{x}_i$ is the original n-dimensional vector. With each $\mathbf{x}_i$, we shall associate a scalar y_i, such that $y_i = 1$ if $\mathbf{x}_i \in \omega_1$ and $y_i = -1$ if $\mathbf{x}_i \in \omega_2$. A given separating hyperplane consists of the points $\mathbf{x}$ such that $\mathbf{w}^T\mathbf{x} + b = 0$. Define d_1 and d_2 to be the distances from the hyperplane to the closest training vector in class ω_1 and ω_2, respectively. We define the *margin* m to be $d_1 + d_2$. The requirement that there is a nonzero (equal) margin on either side of the hyperplane can be formulated as $\mathbf{w}^T\mathbf{x}_i + b \geq 1$ for $\mathbf{x}_i \in \omega_1$ and $\mathbf{w}^T\mathbf{x}_i + b \leq -1$ for $\mathbf{x}_i \in \omega_2$, or

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \ i = 1, \ldots, N. \tag{2.7}$$

For any point (on either side of the hyperplane) for which (2.7) achieves equality, the distance from that point to the hyperplane is $1/\|\mathbf{w}\|$, so that the margin is equal to $2/\|\mathbf{w}\|$. The maximum margin hyperplane is then found by minimizing $\|\mathbf{w}\|$ (or $\|\mathbf{w}\|^2$) subject to (2.7).

The standard approach to constrained minimization problems is to use Lagrange multipliers, which leads to the *primal form* of the objective function

$$L_P = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1] \tag{2.8}$$

where $\alpha_i \geq 0, i = 1, \ldots, N$, are the Lagrange multipliers. We seek to minimize L_P with respect to $\mathbf{w}$ and b, and maximize it with respect to α_i, subject to $\alpha_i \geq 0$. The conditions

$\partial L_P/\partial \mathbf{w} = 0$ and $\partial L_P/\partial b = 0$ become

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i, \tag{2.9}$$

$$0 = \sum_{i=1}^{N} \alpha_i y_i. \tag{2.10}$$

Substitution of (2.9) and (2.10) into equation (2.8) yields the *dual form* of the Lagrangian

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j, \tag{2.11}$$

which we maximize subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i y_i = 0$.

The dual formulation makes the problem of finding the Lagrange multipliers α_i more manageable, but finding the solution usually requires a computer. Once these parameters are found, we can use equation (2.9) to find $\mathbf{w}$. The value for b can be obtained from the Karush-Kuhn-Tucker (KKT) conditions, which are necessary and sufficient conditions for an optimization problem with inequality and equality constraints. The KKT conditions for the primary formulation (2.8) are

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0, \qquad \frac{\partial L_P}{\partial b} = 0,$$

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 \geq 0, \ i = 1, \ldots, N, \qquad \alpha_i \geq 0, \ i = 1, \ldots, N,$$

$$\alpha_i[y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1] = 0, \ i = 1, \ldots, N.$$

Note that (2.9) and (2.10) correspond to the first two. In particular, the last condition $\alpha_i[y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1] = 0$ is known as the KKT complementarity condition, and it distinguishes between *active* and *inactive* constraints. The active constraints correspond to $y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1$, so that $\alpha_i \geq 0$. These $\mathbf{x}_i$'s are precisely those which define the margin. When $y_i(\mathbf{w}^T \mathbf{x}_i + b) > 1$, the vector $\mathbf{x}_i$ is away from the margin, and $\alpha_i = 0$, so the constraint is

inactive. To find b, we can select a vector $\mathbf{x}_i$ corresponding to an active constraint, so that $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$, where $\mathbf{w}$ is given by (2.9). However, to minimize roundoff error and for a simpler decision rule, b would be averaged over all the active constraints.

The set of vectors $\mathbf{x}_i$ with active constraints are called *support vectors*, and they are the ones for which equality holds in (2.7). They alone determine the values of $\mathbf{w}$ and b. If any of the other training samples were moved (provided they do not cross over into the margin) or deleted, it would not affect the solution for the separating hyperplane. If we denote the set of support vectors by $\mathcal{SV}$, then this fact becomes obvious by noting that

$$\mathbf{w} = \sum_{i \in \mathcal{SV}} \alpha_i y_i \mathbf{x}_i, \tag{2.12}$$

$$\begin{aligned} b &= \frac{1}{N_{\mathcal{SV}}} \left[\sum_{i \in \mathcal{SV}} y_i - \mathbf{w}^T \sum_{i \in \mathcal{SV}} \mathbf{x}_i \right], \\ &= \frac{1}{N_{\mathcal{SV}}} \left[\sum_{i \in \mathcal{SV}} y_i - \sum_{i \in \mathcal{SV}} \sum_{j \in \mathcal{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x}_j \right] \end{aligned} \tag{2.13}$$

where $N_{\mathcal{SV}}$ is the number of support vectors. The actual decision rule for a new vector $\mathbf{x}$ is given by the sign of $\mathbf{w}^T\mathbf{x} + b$. Plugging in (2.12) and (2.13) gives us the rule: assign $\mathbf{x}$ to ω_1 if and only if

$$\sum_{i \in \mathcal{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + \frac{1}{N_{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} y_i - \frac{1}{N_{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} \sum_{j \in \mathcal{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x}_j > 0.$$

We saw how any linear decision rule can be generalized by mapping the original space $\mathbb{R}^n$ into $\mathbb{R}^{\hat{n}}$ by some (possibly nonlinear) function $\mathbf{y}(\mathbf{x})$. This sort of generalization for the support vector machines is accomplished by a simple trick. Note that in the dual formulation (2.11), the data appears only in the forms of the inner products $\mathbf{x}_i^T\mathbf{x}_j$. So if we first mapped data into $\mathbb{R}^{\hat{n}}$, we would still use the data only in the form of inner products $\mathbf{y}_i^T\mathbf{y}_j$. Hence, if there were a "kernel function" K such that $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{y}_i^T\mathbf{y}_j$, then we would never even

need to know explicitly which mapping $\mathbf{y} : \mathbb{R}^n \to \mathbb{R}^{\hat{n}}$ was chosen. The SVM training (finding α_i) and decision rule would remain almost unchanged, except that all of the inner products would be replaced by the values of the kernel function K. Popular kernel functions include $K(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1^T \mathbf{x}_2 + 1)^p$, which results in a classifier that is a polynomial of degree p in the data, and $K(\mathbf{x}_1, \mathbf{x}_2) = e^{\|\mathbf{x}_1 - \mathbf{x}_2\|^2/2\sigma^2}$, which gives a Gaussian radial basis function classifier, where the image space is infinite dimensional.

The above derivation works only if the training data is separable. To apply SVMs to non-separable data, we need to introduce "slack" variables $\xi_i,\ i = 1, \ldots, N$, so that constraints (2.7) become

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i,\ i = 1, \ldots, N, \tag{2.14}$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, N. \tag{2.15}$$

Thus $\xi_i = 0$ if $\mathbf{x}_i$ is correctly classified, $0 < \xi_i < 1$ if $\mathbf{x}_i$ is inside the margin, and $\xi_i > 1$ if $\mathbf{x}_i$ is misclassified. Therefore $\sum_{i=1}^N \xi_i$ can be used as an extra cost term in the primary formulation. Hence we minimize $\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^N \xi_i$ subject to (2.14) and (2.15). The primary form of the Lagrangian becomes

$$L_P = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^N \beta_i \xi_i,$$

where $\alpha_i \geq 0$ and $\beta_i \geq 0$ are the Lagrange multipliers. Differentiating with respect to $\mathbf{w}$, b, and ξ_i, we obtain

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i,$$

$$0 = \sum_{i=1}^N \alpha_i y_i,$$

$$C - \alpha_i - \beta_i = 0,\ i = 1, \ldots, N,$$

where the last equation implies that α_i and β_i are each bounded by C. Substituting these into L_P we obtain

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j.$$

This is maximized with respect to α_i, subject to constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0,$$
$$0 \leq \alpha_i \leq C, \ i = 1, \ldots, N.$$

Note that the only difference in L_D from the separable case is the upper bound on α_i. The KKT complementarity conditions become

$$\alpha_i [y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 + \xi_i] = 0,$$

$$\beta_i \xi_i = (C - \alpha_i)\xi_i = 0.$$

Vectors for which $\alpha_i > 0$ are the support vectors, because this implies that $\xi_i = 0$, so these vectors lie on the margin. Nonzero slack variables only occur when $\alpha_i = C$. The vector $\mathbf{w}$ is obtained by setting $\partial L_P / \partial \mathbf{w}$ equal to zero, and b by choosing one of (or averaging over) the samples for which $0 < \alpha_i < C$.

2.3 Dimensionality Considerations

In this paper, we propose a method of classification which can be used when the dimension of the feature vectors is much larger than the number of training samples, i.e. $n >> N$. In situations like this, many of the existing classification techniques fall short. The decision theoretic approach, for example, requires knowledge of the class-conditional pdfs. If there is no prior information about their form, then they must be estimated non-parametrically. However, this is not possible unless $N >> n$, because the number of data points needed

grows exponentially with n. If we assume some parametric form of the class-conditional pdfs (usually normal), then we still have to estimate the $O(n^2)$ parameters in the covariance matrix Σ. The maximum likelihood estimate would only have rank N, so Σ would not even be invertible. If we use a hyperplane to separate the data (which is always possible in this situation), we still have many degrees of freedom, and it is possible to overfit the data.

2.3.1 Principle Component Analysis

One way to deal with a high-dimensional feature space is to combine features. Principle Component Analysis seeks to project high-dimensional data into a lower dimensional space that best represents it in the least-squares sense. Consider the problem of representing the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^n$ by a single vector $\mathbf{x}_0$ such that the sum of the squares of the distances from $\mathbf{x}_i$ to $\mathbf{x}_0$ is as small as possible, i.e. $J_0(\mathbf{x}_0) = \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{x}_0\|^2$ is minimized. It is easy to show that $\mathbf{x}_0 = \mathbf{m}$, where $\mathbf{m}$ is the mean of the N vectors. If we think of each $\mathbf{x}_i$ as being projected onto $\mathbf{x}_0$, then we have a zero-dimensional representation of all the data. To find a one-dimensional representation, let $\mathbf{e}$ be a unit vector in a direction of our choice. We wish to project each $\mathbf{x}_i$ onto the line passing through $\mathbf{m}$ in the direction of $\mathbf{e}$. This is done by choosing the parameters a_i to minimize

$$\begin{aligned} J_1(a_1, \ldots, a_N, \mathbf{e}) &= \sum_{i=1}^N \|\mathbf{x}_i - (\mathbf{m} + a_i\mathbf{e})\|^2, \\ &= \sum_{i=1}^N \left(a_i^2 - 2a_i\mathbf{e}^T(\mathbf{x}_i - \mathbf{m}) + \|\mathbf{x}_i - \mathbf{m}\|^2\right). \end{aligned}$$

Setting $\frac{\partial J_1}{\partial a_i} = 0$ gives $a_i = \mathbf{e}^T(\mathbf{x}_i - \mathbf{m})$. Plugging this into J_1, we obtain

$$\begin{aligned} J_1(a_1, \ldots, a_N, \mathbf{e}) &= \sum_{i=1}^N \left(-\mathbf{e}^T(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T\mathbf{e} + \|\mathbf{x}_i - \mathbf{m}\|^2\right), \\ &= -\mathbf{e}^T\mathbf{S}\mathbf{e} + \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{m}\|^2, \end{aligned}$$

where $\mathbf{S} = \sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T$ is the scatter matrix. The vector $\mathbf{e}$ that minimizes J_1 must maximize $\mathbf{e}^T\mathbf{S}\mathbf{e}$, subject to $\|\mathbf{e}\| = 1$. Therefore, we maximize $L = \mathbf{e}^T\mathbf{S}\mathbf{e} - \lambda(\mathbf{e}^T\mathbf{e} - 1)$, where λ is the Lagrange multiplier. Setting $\frac{\partial L}{\partial \mathbf{e}} = 0$ gives $2\mathbf{S}\mathbf{e} - 2\lambda\mathbf{e} = 0$, or

$$\mathbf{S}\mathbf{e} = \lambda\mathbf{e},$$

so that $\mathbf{e}$ must be an eigenvector of the scatter matrix. In particular, since $\mathbf{e}^T\mathbf{S}\mathbf{e} = \lambda\mathbf{e}^T\mathbf{e} = \lambda$, we must choose the eigenvector corresponding to the largest eigenvalue.

This result can also be obtained for a $\hat{n}$-dimensional projection, where $\hat{n} < n$. In this case, we seek to minimize

$$J_{\hat{n}} = \sum_{i=1}^{N} \|\mathbf{x}_i - \left(\mathbf{m} + \sum_{j=1}^{\hat{n}} a_{i,j}\mathbf{e}_j\right)\|^2.$$

It can similarly be shown that the best choices for $\mathbf{e}_j$ are the directions of the eigenvectors of $\mathbf{S}$ corresponding to the $\hat{n}$ largest eigenvalues. Geometrically, if we can think of $\mathbf{x}_i, i = 1, \ldots, N$, as occupying a ellipsoidal cloud, then the best $\hat{n}$ dimensional representation (in the least squares sense) of the data is to project it onto the principle axes of the ellipsoid. The coefficients $a_{i,j}$ are called the *principle components* of $\mathbf{x}_i$.

2.3.2 Discriminant Analysis

While Principle Component Analysis deals with projecting a group of vectors onto a subspace that best represents them as a group (in the least squares sense), the goal of Discriminant Analysis is to project the data onto the subspace that best separates two groups (also in the least squares sense). The is exactly the idea behind linear discriminant functions: to project the n-dimensional vectors onto a one-dimensional subspace so that the two groups are separated well. We have already seen a connection between the Least Squared Error

criterion function and Fisher's Linear Discriminant. We will now see the motivation behind (1.1). Suppose we wish to separate $\{\mathbf{x}_{1,j},\ j = 1, \ldots, N_1\}$ from $\{\mathbf{x}_{2,j},\ j = 1, \ldots, N_2\}$. Then we need to choose $\mathbf{w}$ so that the scalar products $y_{i,j} = \mathbf{w}^T\mathbf{x}_{i,j}$ have $\{y_{1,j}\}$ well separated from $\{y_{2,j}\}$. Geometrically, $y_{i,j}$ represents the projection of $\mathbf{x}_{i,j}$ onto $\mathbf{w}$ (assuming $\|\mathbf{w}\| = 1$, although the magnitude of $\mathbf{w}$ is unimportant). A measure of the separation between the projected points is the square of the difference of the sample means (m_i):

$$\begin{aligned}
(m_1 - m_2)^2 &= \left[\frac{1}{N_1}\sum_{j=1}^{N_1} y_{1,j} - \frac{1}{N_2}\sum_{j=1}^{N_2} y_{2,j}\right]^2, \\
&= \left[\frac{1}{N_1}\sum_{j=1}^{N_1} \mathbf{w}^T\mathbf{x}_{1,j} - \frac{1}{N_2}\sum_{j=1}^{N_2} \mathbf{w}^T\mathbf{x}_{2,j}\right]^2, \\
&= \left[\mathbf{w}^T(\mathbf{m}_1 - \mathbf{m}_2)\right]^2, \\
&= \mathbf{w}^T\mathbf{S}_B\mathbf{w},
\end{aligned}$$

where $\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T$ is called the *between-class scatter matrix*. We would like the difference between the means to be large relative to the variances, so we define the scatter for projected samples from class ω_i to be

$$\begin{aligned}
s_i^2 &= \sum_{j=1}^{N_i}(y_{i,j} - m_i)^2, \\
&= \sum_{j=1}^{N_i}(\mathbf{w}^T\mathbf{x}_{i,j} - \mathbf{w}^T\mathbf{m}_i)^2, \\
&= \mathbf{w}^T\left[\sum_{j=1}^{N_i}(\mathbf{x}_{i,j} - \mathbf{m}_i)(\mathbf{x}_{i,j} - \mathbf{m}_i)^T\right]\mathbf{w}, \\
&= \mathbf{w}^T\mathbf{S}_i\mathbf{w}.
\end{aligned}$$

where $\mathbf{S}_i$ is the scatter matrix for class ω_i. The sum of these scatters can be written as $s_1^2 + s_2^2 = \mathbf{w}^T(\mathbf{S}_1 + \mathbf{S}_2)\mathbf{w} = \mathbf{w}^T\mathbf{S}_W\mathbf{w}$, where $\mathbf{S}_W$ is the *within-class scatter matrix*. Fisher's

linear discriminant is the vector $\mathbf{w}$ maximizing the ratio

$$\begin{aligned} J(\mathbf{w}) &= \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}, \\ &= \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}, \end{aligned}$$

which coincides with (1.1).

2.4 Matrix-Variate Normal Distribution

Let us assume that, for class ω_i, the class-conditional density $p(\mathbf{x}|\omega_i)$ is multivariate normal. The training process would typically involve estimating the mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. While this model works well for many applications (ignoring dimensionality considerations), it is limited because it can only capture relationships within the feature vector components (via the covariance matrix $\boldsymbol{\Sigma}_i$). What if there are relationships not only within these components, but also among vectors of the same class? And how would one capture these relationships?

With these questions as motivation, we consider matrix-variate normal prior class-conditional densities, which are a generalization of multivariate normal priors because it allows for correlations between components of different vectors of the same class. We have not encountered this in any of the literature except for classification on the basis of repeated measurements [Choi, 1972, Gupta, 1986].

We now present definitions and some useful results for matrix-variate normal random variables. Since they are defined in terms of multivariate normal variables, we will present a more direct proof and forego some of the theory. The reader is directed to Gupta and Nagar [2000] for a more elegant exposition.

Definition 2.4.1. Let $\mathbf{A}$ be an $m \times n$ matrix and $\mathbf{B}$ be $p \times q$. Then we define the Kronecker product $\mathbf{A} \otimes \mathbf{B}$ to be the $mp \times nq$ matrix which can be written in block form as

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{pmatrix}$$

Definition 2.4.2. Let $\mathbf{X}$ be the $m \times n$ matrix $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$, where $\mathbf{x}_i \in \mathbb{R}^m,\ i = 1, \ldots, n$. Then we define $\text{vec}(\mathbf{X})$ to be the $mn \times 1$ matrix

$$\text{vec}(\mathbf{X}) = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}$$

Definition 2.4.3. The random $p \times n$ matrix $\mathbf{X}$ is said to have a matrix-variate normal distribution with mean matrix $\mathbf{M}$ $(p \times n)$ and covariance matrix $\mathbf{\Sigma} \otimes \mathbf{\Psi}$, where $\mathbf{\Sigma}$ $(p \times p)$ and $\mathbf{\Psi}$ $(n \times n)$ are positive definite, if $\text{vec}(\mathbf{X}^T) \sim N_{pn}(\text{vec}(\mathbf{M}^T), \mathbf{\Sigma} \otimes \mathbf{\Psi})$. We will use the notation

$$\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Psi}).$$

It can be shown that the pdf of $\mathbf{X}$ is

$$f(\mathbf{X}) = (2\pi)^{-\frac{np}{2}} |\mathbf{\Sigma}|^{-\frac{n}{2}} |\mathbf{\Psi}|^{-\frac{p}{2}} \text{etr}\left\{-\frac{1}{2}\mathbf{\Sigma}^{-1}(\mathbf{X} - \mathbf{M})\mathbf{\Psi}^{-1}(\mathbf{X} - \mathbf{M})^T\right\}.$$

Theorem 2.4.4. *Suppose* $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Psi})$. *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *and* $p \times p$ *matrices, respectively. Then*

$$\text{E}((\mathbf{X} - \mathbf{M})\mathbf{A}(\mathbf{X} - \mathbf{M})^T) = \text{tr}(\mathbf{\Psi}\mathbf{A})\mathbf{\Sigma},$$

$$\text{E}((\mathbf{X} - \mathbf{M})^T\mathbf{B}(\mathbf{X} - \mathbf{M})) = \text{tr}(\mathbf{\Sigma}\mathbf{B})\mathbf{\Psi},$$

Proof. Without loss of generality, let us assume $\mathbf{M} = \mathbf{0}$. If we write $\mathbf{y} = \text{vec}(\mathbf{X}^T)$, then by definition we have $\text{E}(\mathbf{y}\mathbf{y}^T) = \mathbf{\Sigma} \otimes \mathbf{\Psi}$. Identifying each component tells us that

$$\text{E}(x_{i,j}x_{k,l}) = \sigma_{i,k}\psi_{j,l},\ 1 \leq i, k \leq p,\ 1 \leq j, l \leq n.$$

To prove the first, note that the $(k,l)^{th}$ component of $\mathbf{XAX}^T$ is $\sum_{s=1}^n x_{k,s} \sum_{r=1}^n a_{s,r}x_{l,r}$, for $1 \leq k,l \leq p$. Hence, we have

$$\begin{aligned}
\left(\mathrm{E}(\mathbf{XAX}^T)\right)_{k,l} &= \mathrm{E}\left(\sum_{s=1}^n x_{k,s} \sum_{r=1}^n a_{s,r}x_{l,r}\right), \\
&= \sum_{s=1}^n \sum_{r=1}^n a_{s,r}\mathrm{E}(x_{k,s}x_{l,r}), \\
&= \sum_{s=1}^n \sum_{r=1}^n a_{s,r}\sigma_{k,l}\psi_{s,r}, \\
&= \sigma_{k,l} \sum_{s=1}^n \sum_{r=1}^n a_{s,r}\psi_{s,r}, \\
&= \sigma_{k,l}\ \mathrm{tr}(\mathbf{\Psi A}).
\end{aligned}$$

For the second, we have for $1 \leq k,l \leq n$,

$$\begin{aligned}
\left(\mathrm{E}(\mathbf{X}^T\mathbf{BX})\right)_{k,l} &= \mathrm{E}\left(\sum_{s=1}^p x_{s,k} \sum_{r=1}^p b_{s,r}x_{r,l}\right), \\
&= \sum_{s=1}^p \sum_{r=1}^p b_{s,r}\mathrm{E}(x_{s,k}x_{r,l}), \\
&= \sum_{s=1}^p \sum_{r=1}^p b_{s,r}\sigma_{s,r}\psi_{k,l}, \\
&= \psi_{k,l} \sum_{s=1}^p \sum_{r=1}^p b_{s,r}\sigma_{s,r}, \\
&= \psi_{k,l}\ \mathrm{tr}(\mathbf{\Sigma B}).
\end{aligned}$$

□

Theorem 2.4.5. *Let* $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Psi})$. *Let* $\mathbf{D}$ *be and* $m \times p$ *matrix of rank* $m \leq p$, *and let* $\mathbf{C}$ *be* $n \times t$ *with rank* $t \leq n$. *Then*

$$\mathbf{DXC} \sim N_{m,t}\left(\mathbf{DMC}, (\mathbf{D\Sigma D}^T) \otimes (\mathbf{C}^T\mathbf{\Psi C})\right).$$

Corollary 2.4.6. *Let* $\mathbf{X} \sim N_{p,n}(\mathbf{M}, \mathbf{\Sigma} \otimes \mathbf{\Psi})$ *and let* $\mathbf{x}_i$ *and* $\mathbf{m}_i$ *be the* i^{th} *columns of* $\mathbf{X}$ *and* $\mathbf{M}$*, respectively. Then*

$$\mathbf{x}_i \sim N(\mathbf{m}_i, \psi_{i,i}\mathbf{\Sigma}),\ i = 1, \ldots, n.$$

Proof. In Theorem 2.4.5, let $\mathbf{D}$ be the identity matrix and $\mathbf{C}$ the column vector with a 1 in the i^{th} place and zeros everywhere else. □

3

CLASSIFICATION RULE BASED ON MATRIX DISTRIBUTIONS

3.1 Theory

In this chapter, we propose a new C-class classification rule based on training samples $\mathbf{x}_{i,j},\ j = 1,\ldots,N_i,\ i = 1,\ldots,C$, which is motivated by situations where the length of vectors n is much larger than the total number of training samples $N = \sum_{i=1}^{C} N_i$. We will develop the theory behind the classification rule and propose several implementations – based on different prior distributions $p(\mathbf{a}|\omega_i)$. We will then outline the calculations for each method.

The most common decision theoretic approach is to assume that the vectors from class ω_i obey a multivariate normal distribution, i.e.

$$\mathbf{x}_{i,j} \sim N(\mathbf{m}_i, \mathbf{\Sigma}_i),\ j = 1,\ldots,N_i,\ i = 1,\ldots,C. \tag{1.1}$$

Under this assumption, the covariance matrix $\mathbf{\Sigma}_i$ reflects relationships between components of a vector from class ω_i, but vectors from different classes are assumed to be independent. However, in certain situations, there may be relationships between vectors of different classes. For example, if $\mathbf{x}_{i,j}$ are images of a certain human organ with or without a disease, then we would expect the images from the two classes to be related. To model these kinds of

relationships, we make the additional assumption that the $n \times N$ matrix

$$\mathbf{X} = [\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{1,N_1}, \mathbf{x}_{2,1}, \ldots, \mathbf{x}_{2,N_2}, \ldots, \mathbf{x}_{C,1}, \ldots, \mathbf{x}_{C,N_C}] \tag{1.2}$$

came from a matrix-variate normal distribution

$$\mathbf{X} \sim N(\boldsymbol{\Theta}, \mathbf{V} \otimes \boldsymbol{\Psi}),$$

for some $\boldsymbol{\Theta}$, $\mathbf{V}$ and $\boldsymbol{\Psi}$. For notational convenience, we will define

$$s_{i,j} = N_1 + \ldots + N_{i-1} + j,$$

so that $\mathbf{x}_{i,j}$ is the $s_{i,j}^{th}$ column of $\mathbf{X}$. Further, let us denote the canonical vectors in $\mathbb{R}^N$ as

$$\boldsymbol{\nu}_i = (\underbrace{0, \ldots, 0}_{i-1}, 1, \underbrace{0, \ldots, 0}_{N-i})^T, \ i = 1, \ldots, N,$$

and define the vectors $\mathbf{e}_i \in \mathbb{R}^N, \ i = 1, \ldots, C$, such that $\mathbf{e}_i$ has ones in the places corresponding to class ω_i (see (1.2) and zeros elsewhere. That is,

$$\begin{aligned} \mathbf{e}_i &= \sum_{j=s_{i1}}^{s_{iN_i}} \boldsymbol{\nu}_j \\ &= (0, \ldots, 0, \underbrace{1, \ldots, 1}_{s_{i1}, \ldots, s_{iN_i}}, 0, \ldots, 0)^T. \end{aligned}$$

(Note: we will use $\boldsymbol{\nu}_i$ to denote canonical basis vectors in spaces other than $\mathbb{R}^N$ when there is no ambiguity in the dimension.) Finally, we define the matrix $\mathcal{E} \in \mathbb{R}^{N \times C}$ to be

$$\mathcal{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_C]. \tag{1.3}$$

By Corollary 2.4.6 we have

$$\mathbf{x}_{i,j} \sim N(\boldsymbol{\theta}_{s_{ij}}, \psi_{s_{ij},s_{ij}} \mathbf{V}),$$

where $\boldsymbol{\theta}_{s_{ij}}$ is the appropriate column of $\boldsymbol{\Theta}$. In order for this to be consistent with (1.1), we must choose $\boldsymbol{\theta}_{s_{ij}} = \mathbf{m}_i$ and $\psi_{s_{ij},s_{ij}}\mathbf{V} = \boldsymbol{\Sigma}_i$. In particular, we will assume that

$$\boldsymbol{\Sigma}_i = \sigma_i^2 \boldsymbol{\Sigma}, \tag{1.4}$$

for some common covariance matrix $\boldsymbol{\Sigma}$. Hence we set $\mathbf{V} = \boldsymbol{\Sigma}$ and $\psi_{s_{ij},s_{ij}} = \sigma_i^2$, obtaining

$$\mathbf{X} \sim N(\mathbf{M}\mathcal{E}^T, \boldsymbol{\Sigma} \otimes \boldsymbol{\Psi}), \tag{1.5}$$

where

$$\begin{aligned}
\mathbf{M} &= [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_C], \\
\mathbf{M}\mathcal{E}^T &= [\mathbf{m}_1, \ldots, \mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_2, \ldots, \mathbf{m}_C, \ldots, \mathbf{m}_C], \\
\operatorname{diag}(\boldsymbol{\Psi}) &= (\sigma_1^2, \ldots, \sigma_1^2, \sigma_2^2, \ldots, \sigma_2^2, \ldots, \sigma_C^2, \ldots, \sigma_C^2).
\end{aligned}$$

In what follows, we will assume that the vectors $\{\mathbf{m}_1, \ldots, \mathbf{m}_C\}$ are linearly independent.

We now wish to classify a new vector $\mathbf{z} \in \mathbb{R}^n$. We would like to write $\mathbf{z}$ as a linear combination of the training samples, but this will not be possible unless $\mathbf{z} \in L(\mathbf{X}) \equiv \operatorname{span}\{\mathbf{x}_{i,j}\}$. Hence, we will write

$$\mathbf{z} = \sum_{i,j} \mathbf{x}_{i,j} a_{i,j} + \boldsymbol{\delta},$$

where $\boldsymbol{\delta}$ represents the random deviation of $\mathbf{z}$ from $L(\mathbf{X})$. If we introduce the vector

$$\mathbf{a} = (a_{1,1}, \ldots, a_{1,N_1}, a_{2,1}, \ldots, a_{2,N_2}, \ldots, a_{C,1}, \ldots, a_{C,N_C})^T \in \mathbb{R}^N,$$

then this can be written more compactly as

$$\mathbf{z} = \mathbf{X}\mathbf{a} + \boldsymbol{\delta}. \tag{1.6}$$

The coefficient vector $\mathbf{a}$ can be interpreted as the coefficients of the projection of $\mathbf{z}$ onto $L(\mathbf{X})$, and classification based on these coefficients is similar to the idea of a linear SVM. We shall

discuss this relationship in more detail in Section 3.6. The vector $\boldsymbol{\delta}$ can be interpreted as the orthogonal component in $L(\mathbf{X})^{\perp}$. However, since we will be applying the class-conditional priors to the vector $\mathbf{a}$, it will be convenient for the sake of the decision rule to assume that $\boldsymbol{\delta}$ is independent Gaussian noise, i.e.

$$\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{1.7}$$

From (1.6) and (1.7), the pdf of $\mathbf{z}|\mathbf{a}, \mathbf{X}$ is then given by

$$p(\mathbf{z}|\mathbf{a}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{z} - \mathbf{X}\mathbf{a})^T(\mathbf{z} - \mathbf{X}\mathbf{a})\right\}. \tag{1.8}$$

Since we treat the vector $\boldsymbol{\delta}$ as a "deviation", we will classify the new vector $\mathbf{z}$ on the basis of $\mathbf{X}\mathbf{a}$, where we will choose $\mathbf{a}$ so that $\mathbf{z}$ is classified into class ω_i whenever $\mathbf{X}\mathbf{a} \sim N(\mathbf{m}_i, \sigma_i^2 \boldsymbol{\Sigma})$. From the properties of matrix-variate normal distributions, we establish:

Theorem 3.1.1. *For the matrix-variate random variable* $\mathbf{X}$ *given by (1.5), we have*

$$\mathrm{E}(\mathbf{X}\mathbf{a}|\mathbf{a}) = \mathbf{M}\boldsymbol{\mathcal{E}}^T\mathbf{a},$$

$$\mathrm{Cov}(\mathbf{X}\mathbf{a}|\mathbf{a}) = (\mathbf{a}^T\boldsymbol{\Psi}\mathbf{a})\boldsymbol{\Sigma}.$$

Proof. The calculations are straightforward and follow from (1.5) and Theorem 2.4.4,

$$\mathrm{E}(\mathbf{X}\mathbf{a}|\mathbf{a}) = \mathbf{M}\boldsymbol{\mathcal{E}}^T\mathbf{a},$$

$$\begin{aligned}
\mathrm{Cov}(\mathbf{X}\mathbf{a}|\mathbf{a}) &= \mathrm{E}\left[(\mathbf{X}\mathbf{a} - \mathbf{M}\boldsymbol{\mathcal{E}}^T\mathbf{a})(\mathbf{X}\mathbf{a} - \mathbf{M}\boldsymbol{\mathcal{E}}^T\mathbf{a})^T\right] \\
&= \mathrm{E}\left[(\mathbf{X} - \mathbf{M}\boldsymbol{\mathcal{E}}^T)\mathbf{a}\mathbf{a}^T(\mathbf{X} - \mathbf{M}\boldsymbol{\mathcal{E}}^T)^T\right] \\
&= \mathrm{tr}(\boldsymbol{\Psi}\mathbf{a}\mathbf{a}^T)\boldsymbol{\Sigma} \\
&= (\mathbf{a}^T\boldsymbol{\Psi}\mathbf{a})\boldsymbol{\Sigma}.
\end{aligned}$$

□

Recall from (1.1) and (1.4) that vectors from class ω_i are normally distributed with mean $\mathbf{m}_i$ and covariance matrix $\sigma_i^2\mathbf{\Sigma}$. Therefore, $\mathbf{z}$ belongs to class ω_i whenever $\mathrm{E}(\mathbf{Xa}|\mathbf{a}) = \mathbf{m}_i$ and $\mathrm{Cov}(\mathbf{Xa}|\mathbf{a}) = \sigma_i^2\mathbf{\Sigma}$, i.e. if and only if $\mathbf{a}$ satisfies the two relations

$$\mathbf{M}\mathcal{E}^T\mathbf{a} = \mathbf{m}_i, \qquad\qquad (\mathbf{a}^T\mathbf{\Psi}\mathbf{a})\mathbf{\Sigma} = \sigma_i^2\mathbf{\Sigma}.$$

In the first equation, we note that $\mathbf{m}_i$ is the i^{th} column of $\mathbf{M}$, and in the second we can drop $\mathbf{\Sigma}$. Therefore, these equations can be written

$$\mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i, \qquad\qquad \mathbf{a}^T\mathbf{\Psi}\mathbf{a} = \sigma_i^2, \tag{1.9}$$

where $\boldsymbol{\nu}_i$ is the canonical basis vector in $\mathbb{R}^C$, so that $\mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i$ is equivalent to the C equations $\mathbf{e}_k^T\mathbf{a} = \delta_{i,k},\ k = 1, \ldots, C$, where $\delta_{i,k}$ is the Kronecker delta.

Let us consider the second term in (1.9). From (1.5) and the properties of matrix-normal distributions, we have

$$\mathbf{\Psi} = \frac{1}{\mathrm{tr}(\mathbf{\Sigma})}\left[\mathrm{E}(\mathbf{X}^T\mathbf{X}) - \mathrm{E}(\mathbf{X}^T)\mathrm{E}(\mathbf{X})\right].$$

Plugging this in to (1.9), and using $\mathrm{tr}(\mathbf{\Sigma}) = 1$ (see Section 3.5), we obtain

$$\begin{aligned}\sigma_i^2 &= \mathbf{a}^T\mathbf{\Psi}\mathbf{a} \\ &= \mathbf{a}^T[\mathrm{E}(\mathbf{X}^T\mathbf{X}) - \mathrm{E}(\mathbf{X}^T)\mathrm{E}(\mathbf{X})]\mathbf{a} \\ &= \mathbf{a}^T\mathrm{E}(\mathbf{X}^T\mathbf{X})\mathbf{a} - (\mathbf{M}\mathcal{E}^T\mathbf{a})^T(\mathbf{M}\mathcal{E}^T\mathbf{a}).\end{aligned}$$

But for class ω_i, we have $\mathbf{M}\mathcal{E}^T\mathbf{a} = \mathbf{m}_i$. Hence, the second equation in (1.9) becomes

$$\sigma_i^2 = \mathbf{a}^T\mathrm{E}(\mathbf{X}^T\mathbf{X})\mathbf{a} - \|\mathbf{m}_i\|^2.$$

Thus, we have established the following:

Theorem 3.1.2. *For* $\mathbf{z} = \mathbf{X}\mathbf{a} + \boldsymbol{\delta}$*, we have* $\mathrm{E}(\mathbf{X}\mathbf{a}|\mathbf{a}) = \mathbf{m}_i$ *and* $\mathrm{Cov}(\mathbf{X}\mathbf{a}|\mathbf{a}) = \sigma_i^2\boldsymbol{\Sigma}$ *if and only if*

$$\mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i, \tag{1.10}$$

$$\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a} = \kappa_i^2, \tag{1.11}$$

where we define

$$\boldsymbol{\Omega} = \mathrm{E}(\mathbf{X}^T\mathbf{X}), \tag{1.12}$$

$$\kappa_i^2 = \sigma_i^2 + \|\mathbf{m}_i\|^2. \tag{1.13}$$

Note that (1.10) is equivalent to the C *equations* $\mathbf{e}_k^T\mathbf{a} = \delta_{i,k},\ k = 1, \ldots, C.$

In this approach, we classify a new vector $\mathbf{z}$ into class ω_i not on the basis of the relationship between its components (how close $\mathbf{z}$ is to $\mathbf{m}_i$ and $\mathrm{Cov}(\mathbf{z})$ to $\sigma_i^2\boldsymbol{\Sigma}$), but on the basis of its projection onto the linear space formed by the columns of matrix $\mathbf{X}$, i.e. vector $\mathbf{a}$. The advantage of this approach is that vector $\mathbf{a} \in \mathbb{R}^N$ is of much smaller dimension than $\mathbf{z} \in \mathbb{R}^n$. Hence, we avoid the "curse of dimensionality" by applying the class-conditional priors on the small vector $\mathbf{a}$ instead of on the large vector $\mathbf{z}$ (i.e. $p(\mathbf{a}|\omega_i)$ replaces $p(\mathbf{z}|\omega_i)$). From Theorem 3.1.2, we choose these class-conditional priors $p(\mathbf{a}|\omega_i)$ to be consistent with (1.10) and (1.11), and we classify a new vector according to the posterior probability that it belongs to class $\omega_i, i = 1, \ldots, C$.

To compute the posterior probability that an observed vector $\mathbf{z}$ falls into class ω_i, we use Bayes rule and write

$$p(\omega_i|\mathbf{z}, \mathbf{X}) = \frac{p(\omega_i, \mathbf{z}|\mathbf{X})}{p(\mathbf{z}|\mathbf{X})}. \tag{1.14}$$

Denote by $\pi_i \equiv p(\omega_i)$ the prior probability that a new vector $\mathbf{z}$ falls into class ω_i. Then the

numerator of (1.14) can be written as

$$\begin{aligned} p(\omega_i, \mathbf{z}|\mathbf{X}) &= \pi_i \; p(\mathbf{z}|\omega_i, \mathbf{X}) \\ &= \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{z}, \mathbf{a}|\omega_i, \mathbf{X}) \; \mathrm{d}\mathbf{a} \\ &= \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{a}|\omega_i, \mathbf{X}) \; p(\mathbf{z}|\mathbf{a}, \omega_i, \mathbf{X}) \; \mathrm{d}\mathbf{a}. \end{aligned}$$

But since we will choose $p(\mathbf{a}|\omega_i, \mathbf{X}) = p(\mathbf{a}|\omega_i)$ according to conditions (1.10) and (1.11), and since $p(\mathbf{z}|\mathbf{a}, \omega_i, \mathbf{X}) = p(\mathbf{z}|\mathbf{a}, \mathbf{X})$ is given in (1.8), then we have

$$p(\omega_i, \mathbf{z}|\mathbf{X}) = \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{a}|\omega_i) \; p(\mathbf{z}|\mathbf{a}, \mathbf{X}) \; \mathrm{d}\mathbf{a}.$$

Note that the denominator of (1.14) is just the scaling constant

$$p(\mathbf{z}|\mathbf{X}) = \sum_{i=1}^{C} \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{a}|\omega_i) \; p(\mathbf{z}|\mathbf{a}, \mathbf{X}) \; \mathrm{d}\mathbf{a},$$

which is independent of i. Therefore the posteriors $p(\omega_i|\mathbf{z}, \mathbf{X})$ are proportional to $p(\omega_i, \mathbf{z}|\mathbf{X})$, i.e.

$$p(\omega_i|\mathbf{z}, \mathbf{X}) \propto \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{a}|\omega_i) \; p(\mathbf{z}|\mathbf{a}, \mathbf{X}) \; \mathrm{d}\mathbf{a}. \tag{1.15}$$

The (minimum error-rate) decision rule is determined by choosing the class ω_i with the highest posterior probability (1.14) – or equivalently (1.15). We must then consider the problem of how to specify the prior pdfs $p(\mathbf{a}|\omega_i),\ i = 1, \ldots, C$, according to conditions (1.10) and (1.11). Depending on how we interpret these conditions, this can be done in more than one way. In this paper, we will propose several different interpretations of these conditions, each resulting in a different set of priors on $\mathbf{a}|\omega_i$. Naturally, different choices of the priors yield different decision rules.

3.2 Delta Prior

3.2.1 Derivation

One possibility is to require that constraints (1.10) and (1.11) be satisfied with probability one, i.e.

$$P(\mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i) = 1,$$

$$P(\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a} = \kappa_i^2) = 1.$$

In this case, the support of $p(\mathbf{a}|\omega_i)$ must be contained in the intersection of the C hyperplanes $\mathbf{e}_k^T\mathbf{a} = \delta_{i,k},\ k = 1, \ldots, C$, and the ellipsoid $\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a} = \kappa_i^2$. Formally, we can accomplish this by setting

$$p(\mathbf{a}|\omega_i) = C_i\ \delta(\mathcal{E}^T\mathbf{a} - \boldsymbol{\nu}_i)\ \delta(\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a} - \kappa_i^2), \tag{2.16}$$

where C_i is a scaling factor so that $\int_{R^N} p(\mathbf{a}|\omega_i)\ \mathrm{d}\mathbf{a} = 1$, and $\delta(\cdot)$ is the Dirac delta function. Vector arguments to the delta function are interpreted as a termwise product of C delta functions. We shall refer to (2.16) as the Delta prior.

One drawback of this approach is that when we substitute (2.16) into (1.15), the integral becomes difficult to work out analytically. Fortunately, by choosing a convenient representation of the delta function, we can reduce the N-dimensional integral in (1.15) to a more tractable one-dimensional integral.

3.2.2 Decision Rule

The decision rule for the classification of $\mathbf{z}$ is determined by evaluating (1.15) for $i = 1, \ldots, C$. Without loss of generality, we will consider the case $i = 1$. For this and the following sections, we will use several results:

Definition 3.2.1. Let $\mathbf{x} = (x_1, x_2, \ldots, x_N)^T \in \mathbb{R}^N$. Then we will denote by $\mathbf{x}_l$ the vector $(x_1, x_2, \ldots, x_{N-1})^T \in \mathbb{R}^{N-1}$, where the subscript l indicates that the last element has been dropped. Similarly, we will use $\mathbf{x}_f$ to denote $(x_2, x_3, \ldots, x_N)^T \in \mathbb{R}^{N-1}$, where the first element has been dropped.

Lemma 3.2.2. *Let $\mathbf{b}, \mathbf{v} \in \mathbb{R}^N$, $c \in \mathbb{R}$, and let $\mathbf{\Psi}$ be a symmetric $N \times N$ matrix. If, in the quadratic form $\mathbf{x}^T\mathbf{\Psi}\mathbf{x} + 2\mathbf{b}^T\mathbf{x}$, we replace x_N by $\frac{1}{v_N}(c - x_1v_1 - \ldots - x_{N-1}v_{N-1})$ — perhaps as a result of evaluating the integral $\int g(\mathbf{x}^T\mathbf{\Psi}\mathbf{x} + 2\mathbf{b}^T\mathbf{x})\ \delta(\mathbf{v}^T\mathbf{x} - c)\ \mathrm{d}x_N$ — we have*

$$\mathbf{x}^T\mathbf{\Psi}\mathbf{x} + 2\mathbf{b}^T\mathbf{x}\big|_{x_N=\frac{1}{v_N}(c-x_1v_1-\ldots-x_{N-1}v_{N-1})} = \mathbf{x}_l^T\mathbf{\Phi}\mathbf{x}_l + 2\mathbf{x}_l^T\mathbf{w} + C, \tag{2.17}$$

where

$$\begin{aligned}
\Phi_{i,j} &= \Psi_{i,j} - \frac{v_i\Psi_{N,j} + v_j\Psi_{N,i}}{v_N} + \frac{v_iv_j}{v_N^2}\Psi_{N,N},\ i,j = 1,\ldots,N-1,\\
w_i &= \frac{c}{v_N}\left(\Psi_{N,i} - \frac{v_i\Psi_{N,N}}{v_N}\right) + b_i - \frac{b_N}{v_N}v_i,\ i = 1,\ldots,N-1,\\
C &= \frac{\Psi_{N,N}}{v_N^2}c^2 + \frac{2b_N}{v_N}c.
\end{aligned}$$

Furthermore, if $\mathbf{\Psi}$ is positive definite, then $\mathbf{\Phi}$ is positive definite.

In particular, if $\mathbf{\Psi}$ is the identity matrix $\mathbf{I}_N$, $\mathbf{b} = \mathbf{0}$, and $v_N = 1$, then we have the following relations

$$\mathbf{\Phi} = \mathbf{I}_{N-1} + \mathbf{v}_l\mathbf{v}_l^T, \tag{2.18}$$

$$\mathbf{w} = -c\mathbf{v}_l, \tag{2.19}$$

$$\|\mathbf{v}\|^2 = 1 + \|\mathbf{v}_l\|^2, \tag{2.20}$$

$$\mathbf{\Phi}^{-1} = \mathbf{I}_{N-1} - \frac{1}{|\mathbf{\Phi}|}\mathbf{v}_l\mathbf{v}_l^T, \tag{2.21}$$

$$|\mathbf{\Phi}| = \|\mathbf{v}\|^2. \tag{2.22}$$

The proof of this and other statements are placed in Section 3.7

Lemma 3.2.3. *If the* $\mathbf{A}$ *is positive definite, then*

$$\int \cdots \int_{\mathbb{R}^N} \exp\left\{-\frac{1}{2}\left(\mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c\right)\right\}\, \mathrm{d}\mathbf{x} = \left[\frac{(2\pi)^N \exp\left\{\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} - c\right\}}{|\mathbf{A}|}\right]^{1/2}.$$

Lemma 3.2.4. *If* $\mathbf{\Psi}$ *is positive definite, then*

$$\int \cdots \int_{\mathbb{R}^N} e^{-\mathbf{x}^T\mathbf{\Psi}\mathbf{x}}\, \delta(\mathbf{v}^T\mathbf{x} - c)\, \mathrm{d}\mathbf{x} = \frac{\sqrt{\pi}^{\,N-1}}{|\mathbf{\Psi}|^{1/2}\sqrt{\mathbf{v}^T\mathbf{\Psi}^{-1}\mathbf{v}}}\, \exp\left\{\frac{-c^2}{\mathbf{v}^T\mathbf{\Psi}^{-1}\mathbf{v}}\right\}.$$

Lemma 3.2.5. *Let* $\mathbf{v} = (v_1, \ldots, v_N)^T$, $c > 0$, *and* $\mathbf{d} = (d_1, \ldots, d_N)^T$ *where* $d_i > 0, i = 1, \ldots, N$. *Let* $\mathbf{D} = \mathrm{diag}\{\mathbf{d}\}$ *and define the function*

$$J \equiv J(\mathbf{d}, \mathbf{v}, c) = \int \cdots \int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\mathbf{D}\mathbf{x} + 2\mathbf{v}^T\mathbf{x})}\, \delta(\mathbf{x}^T\mathbf{x} - c)\, \mathrm{d}\mathbf{x}.$$

Then

$$J = \pi^{\frac{N-2}{2}}\, e^{\sum_{j=1}^N \frac{v_j^2}{d_j}} \int_0^\infty \left(\prod_{j=1}^N (d_j^2 + u^2)^{-1/4}\right) e^{\sum_{j=1}^N \frac{-v_j^2 u^2}{d_j(d_j^2+u^2)}} \times \cos\left(-cu + \sum_{j=1}^N \frac{v_j^2 u}{d_j^2 + u^2} + \frac{1}{2}\tan^{-1}\frac{u}{d_j}\right) \mathrm{d}u.$$

To find $p(\omega_1|\mathbf{z}, \mathbf{X})$, we plug (2.16) and (1.8) into (1.15), giving

$$p(\omega_1|\mathbf{z}, \mathbf{X}) \propto \frac{C_1\pi_1}{(2\pi\sigma^2)^{n/2}} \int \cdots \int_{\mathbb{R}^N} \delta(\mathbf{e}_1^T\mathbf{a} - 1)\, \delta(\mathbf{e}_2^T\mathbf{a})\, \delta(\mathbf{a}^T\mathbf{\Omega}\mathbf{a} - \kappa_1^2) \times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{z} - \mathbf{X}\mathbf{a})^T(\mathbf{z} - \mathbf{X}\mathbf{a})\right\}\, \mathrm{d}\mathbf{a}. \tag{2.23}$$

Before we evaluate this integral, we compute C_1. For notational convenience, we will use

$\hat{\mathbf{e}} = \mathbf{e}_1$ and $\hat{\hat{\mathbf{e}}} = \mathbf{e}_2$. From (2.16), we have

$$\begin{aligned}\frac{1}{C_1} &= \int \cdots \int_{\mathbb{R}^N} \delta(\hat{\mathbf{e}}^T\mathbf{a} - 1)\ \delta(\mathbf{a}^T\mathbf{\Omega}\mathbf{a} - \kappa_1^2)\ \delta(\hat{\hat{\mathbf{e}}}^T\mathbf{a})\ \mathrm{d}\mathbf{a} \\ &= \int_{\mathbb{R}} \left[\int \cdots \int_{\mathbb{R}^{N-1}} \delta(\hat{\mathbf{e}}^T\mathbf{a} - 1)\ \delta(\mathbf{a}^T\mathbf{\Omega}\mathbf{a} - \kappa_1^2)\ \delta(\hat{\hat{\mathbf{e}}}^T\mathbf{a})\ \mathrm{d}\mathbf{a}_l \right] \mathrm{d}a_N.\end{aligned}$$

By Lemma 3.2.2, this becomes

$$\frac{1}{C_1} = \int \cdots \int_{\mathbb{R}^{N-1}} \delta(\hat{\mathbf{e}}_l^T\mathbf{a}_l - 1)\ \delta(\mathbf{a}_l^T\mathbf{\Phi}\mathbf{a}_l - \kappa_1^2)\ \mathrm{d}\mathbf{a}_l,$$

where

$$\Phi_{i,j} = \Omega_{i,j} - (\hat{\hat{e}}_i\Omega_{N,j} + \hat{\hat{e}}_j\Omega_{N,i}) + \hat{\hat{e}}_i\hat{\hat{e}}_j\Omega_{N,N},\ i,j = 1,\ldots,N-1.$$

We wish to apply Lemma 3.2.2 again, but the last element of $\hat{\mathbf{e}}_l$ is zero. Hence, we will integrate with respect to the first element of $\mathbf{a}_l$, requiring an obvious modification of Lemma 3.2.2. Continuing, we have

$$\begin{aligned}\frac{1}{C_1} &= \int \cdots \int_{\mathbb{R}^{N-1}} \delta(\hat{\mathbf{e}}_l^T\mathbf{a}_l - 1)\ \delta(\mathbf{a}_l^T\mathbf{\Phi}\mathbf{a}_l - \kappa_1^2)\ \mathrm{d}\mathbf{a}_l \\ &= \int_{\mathbb{R}} \left[\int \cdots \int_{\mathbb{R}^{N-2}} \delta(\hat{\mathbf{e}}_l^T\mathbf{a}_l - 1)\ \delta(\mathbf{a}_l^T\mathbf{\Phi}\mathbf{a}_l - \kappa_1^2)\ \mathrm{d}\mathbf{a}_{fl} \right] \mathrm{d}a_1 \\ &= \int \cdots \int_{\mathbb{R}^{N-2}} \delta(\mathbf{a}_{fl}^T\mathbf{\Upsilon}\mathbf{a}_{fl} + 2\mathbf{a}_{fl}^T\mathbf{w} + D - \kappa_1^2)\ \mathrm{d}\mathbf{a}_{fl},\end{aligned}$$

where

$$\Upsilon_{i-1,j-1} = \Phi_{i,j} - (\hat{e}_i\Phi_{1,j} + \hat{e}_j\Phi_{1,i}) + \hat{e}_i\hat{e}_j\Phi_{1,1},\ i,j = 2,\ldots,N-1,$$

$$w_{i-1} = \Phi_{1,i} - \hat{e}_i\Phi_{1,1},\ i = 2,\ldots,N-1,$$

$$D = \Phi_{1,1}.$$

In terms of the original matrix $\mathbf{\Omega}$, we have

$$\begin{aligned}
\Upsilon_{i-1,j-1} &= \Omega_{i,j} - (\hat{e}_i\Omega_{1,j} + \hat{e}_j\Omega_{1,i}) - (\hat{\hat{e}}_i\Omega_{N,j} + \hat{\hat{e}}_j\Omega_{N,i}) + (\hat{e}_i\hat{\hat{e}}_j + \hat{e}_j\hat{\hat{e}}_i)\Omega_{1,N} \\
&\quad + \hat{e}_i\hat{e}_j\kappa_1^2 + \hat{\hat{e}}_i\hat{\hat{e}}_j\kappa_2^2,\ i,j = 2,\ldots,N-1, \\
w_{i-1} &= \Omega_{1,i} - \hat{\hat{e}}_i\Omega_{1,N} - \hat{e}_i\kappa_1^2,\ i = 2,\ldots,N-1, \\
D &= \Omega_{1,1} = \kappa_1^2.
\end{aligned}$$

This gives

$$\begin{aligned}
\frac{1}{C_1} &= \int\cdots\int_{\mathbb{R}^{N-2}} \delta(\mathbf{a}_{fl}^T\mathbf{\Upsilon}\mathbf{a}_{fl} + 2\mathbf{a}_{fl}^T\mathbf{w})\ \mathrm{d}\mathbf{a}_{fl} \\
&= \int\cdots\int_{\mathbb{R}^{N-2}} \delta\left(\left(\mathbf{a}_{fl} + \mathbf{\Upsilon}^{-1}\mathbf{w}\right)^T\mathbf{\Upsilon}\left(\mathbf{a}_{fl} + \mathbf{\Upsilon}^{-1}\mathbf{w}\right) - \mathbf{w}^T\mathbf{\Upsilon}^{-1}\mathbf{w}\right)\ \mathrm{d}\mathbf{a}_{fl},
\end{aligned}$$

where we have completed the square on the last line. Since $\mathbf{\Upsilon}$ is positive definite, we we can write $\mathbf{\Upsilon} = \sqrt{\mathbf{\Upsilon}}^{\,T}\sqrt{\mathbf{\Upsilon}}$, and perform the change of variables $\hat{\mathbf{a}} = \sqrt{\mathbf{\Upsilon}}\ (\mathbf{a}_{fl} + \mathbf{\Upsilon}^{-1}\mathbf{w})$, so that we obtain

$$\frac{1}{C_1} = |\mathbf{\Upsilon}|^{-1/2}\int\cdots\int_{\mathbb{R}^{N-2}} \delta(\hat{\mathbf{a}}^T\hat{\mathbf{a}} - \mathbf{w}^T\mathbf{\Upsilon}^{-1}\mathbf{w})\ \mathrm{d}\hat{\mathbf{a}}.$$

This integral can be found in Gradshteyn and Ryzhik [2000] as the surface area of an $N-2$ dimensional sphere of radius $r = \sqrt{\mathbf{w}^T\mathbf{\Upsilon}^{-1}\mathbf{w}}$, giving

$$C_1 = \frac{|\mathbf{\Upsilon}|^{1/2}\ \Gamma(\frac{N-2}{2})}{2\sqrt{\pi}^{\,N-2}(\mathbf{w}^T\mathbf{\Upsilon}^{-1}\mathbf{w})^{(N-3)/2}}.$$

Having found C_1, we now evaluate (2.23). We first write it as

$$\begin{aligned}
p(\omega_1|\mathbf{z},\mathbf{X}) \propto \frac{C_1\pi_1}{(2\pi\sigma^2)^{n/2}}\int\cdots\int_{\mathbb{R}^N} &\delta(\hat{\mathbf{e}}^T\mathbf{a} - 1)\ \delta(\hat{\hat{\mathbf{e}}}^T\mathbf{a})\ \delta(\mathbf{a}^T\mathbf{\Omega}\mathbf{a} - \kappa_1^2) \\
&\times \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{a}^T\hat{\mathbf{X}}\mathbf{a} - 2\hat{\mathbf{z}}^T\mathbf{a} + \mathbf{z}^T\mathbf{z}\right)\right\}\ \mathrm{d}\mathbf{a},
\end{aligned}$$

where we have used $\hat{\mathbf{X}} = \mathbf{X}^T\mathbf{X}$ and $\hat{\mathbf{z}} = \mathbf{X}^T\mathbf{z}$. As in the calculation of C_1, we apply Lemma

3.2.2 twice to get

$$p(\omega_1|\mathbf{z},\mathbf{X}) \propto \frac{C_1\pi_1}{(2\pi\sigma^2)^{n/2}} \int\cdots\int_{\mathbb{R}^{N-2}} \delta(\mathbf{a}_{fl}^T\boldsymbol{\Upsilon}\mathbf{a}_{fl} + 2\mathbf{a}_{fl}^T\mathbf{w}) \times \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{a}_{fl}^T\boldsymbol{\Delta}\mathbf{a}_{fl} - 2\mathbf{n}^T\mathbf{a}_{fl} + G\right)\right\} \mathrm{d}\mathbf{a}_{fl},$$

where $\boldsymbol{\Upsilon}$ and $\mathbf{w}$ are the same as before, and $\boldsymbol{\Delta}$, $\mathbf{n}$, and G can be found from Lemma 3.2.2, giving

$$\begin{aligned}
\Delta_{i-1,j-1} &= \hat{X}_{i,j} - (\hat{e}_i X_{1,j} + \hat{e}_j X_{1,i}) - (\hat{\hat{e}}_i X_{N,j} + \hat{\hat{e}}_j X_{N,i}) + (\hat{e}_i\hat{\hat{e}}_j + \hat{e}_j\hat{\hat{e}}_i)X_{1,N} \\
&\quad + \hat{e}_i\hat{e}_j X_{1,1} + \hat{\hat{e}}_i\hat{\hat{e}}_j\hat{X}_{N,N},\ i,j = 2,\ldots,N-1, \\
n_{i-1} &= \hat{z}_i + \hat{e}_i\hat{X}_{1,1} + \hat{\hat{e}}_i\hat{X}_{1,N} - \hat{X}_{1,i},\ i = 2,\ldots,N-1, \\
G &= \mathbf{z}^T\mathbf{z} + \hat{X}_{1,1}.
\end{aligned}$$

We then make a similar substitution as before, setting $\mathbf{y} = \sqrt{\boldsymbol{\Upsilon}}\ (\mathbf{a}_{fl} + \boldsymbol{\Upsilon}^{-1}\mathbf{w})$, which yields

$$p(\omega_1|\mathbf{z},\mathbf{X}) \propto \frac{C_1\pi_1}{|\boldsymbol{\Upsilon}|^{1/2}(2\pi\sigma^2)^{n/2}} \int\cdots\int_{\mathbb{R}^{N-2}} \delta(\mathbf{y}^T\mathbf{y} - \mathbf{w}^T\boldsymbol{\Upsilon}^{-1}\mathbf{w}) \times \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y}^T\hat{\boldsymbol{\Delta}}\mathbf{y} - 2\hat{\mathbf{n}}^T\mathbf{y} + \hat{G}\right)\right\} \mathrm{d}\mathbf{y},$$

where

$$\begin{aligned}
\hat{\boldsymbol{\Delta}} &= \sqrt{\boldsymbol{\Upsilon}}^{\ -T}\boldsymbol{\Delta}\sqrt{\boldsymbol{\Upsilon}}^{\ -1}, \\
\hat{\mathbf{n}} &= \sqrt{\boldsymbol{\Upsilon}}^{\ -T}\mathbf{n} + \sqrt{\boldsymbol{\Upsilon}}^{\ -T}\boldsymbol{\Delta}\boldsymbol{\Upsilon}^{-1}\mathbf{w}, \\
\hat{G} &= G + \mathbf{w}^T\boldsymbol{\Upsilon}^{-T}\boldsymbol{\Delta}\boldsymbol{\Upsilon}^{-1}\mathbf{w} + 2\mathbf{n}^T\boldsymbol{\Upsilon}^{-1}\mathbf{w}.
\end{aligned}$$

In the two quadratic forms $\mathbf{y}^T\mathbf{y}$ and $\mathbf{y}^T\hat{\boldsymbol{\Delta}}\mathbf{y}$, both $\mathbf{I}_{N-2}$ and $\hat{\boldsymbol{\Delta}}$ are positive definite. Therefore, there exists matrix that simultaneously diagonalizes both. In other words, there exists a matrix $\mathbf{Q} \in \mathbb{R}^{(N-2)\times(N-2)}$ such that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and $\mathbf{Q}^T\hat{\boldsymbol{\Delta}}\mathbf{Q} = \mathrm{diag}(\hat{\lambda}_1,\ldots,\hat{\lambda}_{N-2}) \equiv \boldsymbol{\Lambda}$. Let

$\mathbf{y} = \mathbf{Q}\mathbf{x}$, then we have

$$p(\omega_1|\mathbf{z}, \mathbf{X}) \propto \frac{|\mathbf{Q}|\, C_1 \pi_1}{|\boldsymbol{\Upsilon}|^{1/2} (2\pi\sigma^2)^{n/2}} \int \cdots \int_{\mathbb{R}^{N-2}} \delta(\mathbf{x}^T\mathbf{x} - \mathbf{w}^T\boldsymbol{\Upsilon}^{-1}\mathbf{w}) \times \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x} - 2\hat{\mathbf{n}}^T\mathbf{Q}\mathbf{x} + \hat{G}\right)\right\} \, \mathrm{d}\mathbf{x}.$$

This integral can be expressed as a one-dimensional integral, according to Lemma 3.2.5. However, the integrand is very complicated and highly oscillatory, so it is difficult to evaluate. For this reason, we are not pursuing the decision rule based on the Delta prior in this paper.

3.3 Maximum Entropy Prior

3.3.1 Derivation

An alternate interpretation is to require that $p(\mathbf{a}|\omega_i)$ satisfy (1.10) and (1.11) on the average, i.e.

$$\mathrm{E}_{\mathbf{a}|\omega_i}[\mathcal{E}^T\mathbf{a}] = \boldsymbol{\nu}_i,$$

$$\mathrm{E}_{\mathbf{a}|\omega_i}[\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a}] = \kappa_i^2.$$

Under these conditions, we seek the prior which introduces as little new information as possible. A useful way of addressing this problem is through the concept of entropy, which measures the amount of uncertainty in a pdf [Berger, 1985]. For continuous random variables, it is common to define the entropy of a pdf f as $\mathrm{Ent}(f) = -\mathrm{E}[\ln f]$. Therefore, we choose $p(\mathbf{a}|\omega_i)$ to be the function $f_i(\mathbf{a})$ which maximizes

$$\mathrm{Ent}(f_i) = -\mathrm{E}[\ln f_i(\mathbf{a})] = -\int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) \ln f_i(\mathbf{a}) \, \mathrm{d}\mathbf{a},$$

subject to the constraints

$$\delta_{i,k} = \int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) \mathbf{e}_k^T \mathbf{a} \, \mathrm{d}\mathbf{a}, \; k = 1, \ldots, C, \tag{3.24}$$

$$\kappa_i^2 = \int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) \mathbf{a}^T \mathbf{\Omega} \mathbf{a} \, \mathrm{d}\mathbf{a}, \tag{3.25}$$

$$1 = \int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) \, \mathrm{d}\mathbf{a}.$$

In other words, we need to maximize the functional

$$F(f_i) = -\int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) [\ln f_i(\mathbf{a})] \, \mathrm{d}\mathbf{a} - \sum_{j=1}^{C} \lambda_j \left[\int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) \mathbf{e}_j^T \mathbf{a} \, \mathrm{d}\mathbf{a} - \delta_{i,j} \right]$$
$$- \gamma \left[\int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) \mathbf{a}^T \mathbf{\Omega} \mathbf{a} \, \mathrm{d}\mathbf{a} - \kappa_i^2 \right] - \rho \left[\int \cdots \int_{\mathbb{R}^N} f_i(\mathbf{a}) \, \mathrm{d}\mathbf{a} - 1 \right],$$

where λ_j, γ, and ρ are the Lagrange multipliers. The Euler-Lagrange equation becomes

$$0 = -1 - \ln f_i(\mathbf{a}) - \sum_{j=1}^{C} \lambda_j \mathbf{e}_j^T \mathbf{a} - \gamma \mathbf{a}^T \mathbf{\Omega} \mathbf{a} - \rho,$$

so that

$$f_i(\mathbf{a}) = \exp \left\{ -\gamma \mathbf{a}^T \mathbf{\Omega} \mathbf{a} - \sum_{j=1}^{C} \lambda_j \mathbf{e}_j^T \mathbf{a} - \rho - 1 \right\}$$
$$= C \exp \left\{ -\frac{1}{2} \left[\mathbf{a} + \frac{\sum_{j=1}^{C} \lambda_j \mathbf{\Omega}^{-1} \mathbf{e}_j}{2\gamma} \right]^T 2\gamma \mathbf{\Omega} \left[\mathbf{a} + \frac{\sum_{j=1}^{C} \lambda_j \mathbf{\Omega}^{-1} \mathbf{e}_j}{2\gamma} \right] \right\},$$

where C is a normalizing constant that does not depend on $\mathbf{a}$. In this form, we recognize that $f_i(\mathbf{a})$ is the pdf of a multivariate normal random variable

$$\mathbf{a} | \omega_i \sim N \left(-\frac{\sum_{j=1}^{C} \lambda_j \mathbf{\Omega}^{-1} \mathbf{e}_j}{2\gamma}, \frac{\mathbf{\Omega}^{-1}}{2\gamma} \right). \tag{3.26}$$

To find λ_j and γ, we use (3.24) and (3.25). From the properties of multivariate normal variables, these constraints become

$$\delta_{i,k} = -\left[\frac{\sum_{j=1}^{C} \lambda_j \mathbf{e}_k^T \mathbf{\Omega}^{-1} \mathbf{e}_j}{2\gamma} \right], \; k = 1, \ldots, C,$$

$$\kappa_i^2 = \mathrm{tr} \left(\frac{1}{2\gamma} \mathbf{\Omega}^{-1} \mathbf{\Omega} \right) + \left[\frac{\sum_{j=1}^{C} \lambda_j \mathbf{\Omega}^{-1} \mathbf{e}_j}{2\gamma} \right]^T \mathbf{\Omega} \left[\frac{\sum_{j=1}^{C} \lambda_j \mathbf{\Omega}^{-1} \mathbf{e}_j}{2\gamma} \right].$$

Let us define the matrices

$$\mathbf{P} = \mathcal{E}^T \mathbf{\Omega}^{-1} \mathcal{E}, \qquad \mathbf{Q} = \mathbf{P}^{-1}, \tag{3.27}$$

so that $p_{k,l} = \mathbf{e}_k^T \mathbf{\Omega}^{-1} \mathbf{e}_l$. We can then write the constraint equations as

$$0 = \sum_{j=1}^{C} \lambda_j p_{k,j}, \ k = 1, \ldots, C, \ k \neq i, \tag{3.28}$$

$$-2\gamma = \sum_{j=1}^{C} \lambda_j p_{i,j}, \tag{3.29}$$

$$4\gamma^2 \kappa_i^2 = 2\gamma N + \sum_{j=1}^{C} \sum_{k=1}^{C} \lambda_j \lambda_k p_{j,k}. \tag{3.30}$$

To solve equations (3.28) – (3.30), we introduce the vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma} \in \mathbb{R}^C$, where

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_C)^T,$$

$$\boldsymbol{\gamma} = (0, \ldots, \underbrace{\gamma}_{i}, \ldots, 0)^T.$$

Equations (3.28) and (3.29) can be written as $\mathbf{P}\boldsymbol{\lambda} = -2\boldsymbol{\gamma}$, so that

$$\boldsymbol{\lambda} = -2\mathbf{P}^{-1}\boldsymbol{\gamma}.$$

Equation (3.30) then becomes

$$\begin{aligned} 0 &= 4\gamma^2\kappa_i^2 - 2\gamma N - \boldsymbol{\lambda}^T \mathbf{P} \boldsymbol{\lambda} \\ &= 4\gamma^2\kappa_i^2 - 2\gamma N - 4\boldsymbol{\gamma}^T \mathbf{P}^{-1} \boldsymbol{\gamma} \\ &= 4\gamma^2\kappa_i^2 - 2\gamma N - 4q_{i,i}\gamma^2 \\ &= 2\gamma \left[2\gamma(\kappa_i^2 - q_{i,i}) - N \right], \end{aligned}$$

so that

$$\begin{aligned} \gamma &= \frac{N}{2(\kappa_i^2 - q_{i,i})}, \\ \lambda_j &= -\frac{N q_{j,i}}{(\kappa_i^2 - q_{i,i})}. \end{aligned}$$

Plugging in γ and λ_j into (3.26) yields

$$\mathbf{a}|\omega_i \sim N\left(\mathbf{\Omega}^{-1}\sum_{j=1}^{C} q_{j,i}\mathbf{e}_j, \frac{(\kappa_i^2 - q_{i,i})}{N}\mathbf{\Omega}^{-1}\right). \tag{3.31}$$

3.3.2 Decision Rule

To compute the decision rule using the Maximum Entropy prior, we only need to evaluate $p(\omega_i|\mathbf{z}, \mathbf{X})$ in (1.15) using $p(\mathbf{a}|\omega_i)$ in (3.31) and $p(\mathbf{z}|\mathbf{a}, \mathbf{X})$ in (1.8). For notational convenience, we will use

$$r_i = \frac{\kappa_i^2 - q_{i,i}}{N}, \qquad \mathbf{u}_i = \mathbf{\Omega}^{-1}\sum_{j=1}^{C} q_{j,i}\mathbf{e}_j,$$

so that

$$\mathbf{a}|\omega_i \sim N(\mathbf{u}_i, r_i\mathbf{\Omega}^{-1}).$$

Plugging in to (1.15) gives

$$\begin{aligned} p(\omega_i|\mathbf{z}, \mathbf{X}) &\propto \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{a}|\omega_i)\ p(\mathbf{z}|\mathbf{a}, \mathbf{X})\ \mathrm{d}\mathbf{a} \\ &= \frac{\pi_i}{(2\pi\sigma^2)^{n/2}(2\pi)^{N/2}\left|r_i\mathbf{\Omega}^{-1}\right|^{1/2}} \int \cdots \int_{\mathbb{R}^N} \exp\left\{-\frac{1}{2r_i}(\mathbf{a}-\mathbf{u}_i)^T\mathbf{\Omega}(\mathbf{a}-\mathbf{u}_i)\right\} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{z}-\mathbf{X}\mathbf{a})^T(\mathbf{z}-\mathbf{X}\mathbf{a})\right\}\ \mathrm{d}\mathbf{a}. \end{aligned}$$

Dropping the terms that are independent of i and rearranging the exponent, we can write

$$p(\omega_i|\mathbf{z}, \mathbf{X}) \propto \frac{\pi_i}{\left|r_i\mathbf{\Omega}^{-1}\right|^{1/2}} \int \cdots \int_{\mathbb{R}^N} \exp\left\{-\frac{1}{2}\left[(\mathbf{a}-\mathbf{v}_i)^T\bar{\mathbf{\Omega}}_i(\mathbf{a}-\mathbf{v}_i) + H\right]\right\}\ \mathrm{d}\mathbf{a},$$

where

$$\begin{aligned} \mathbf{v}_i &= \left(\frac{1}{r_i}\mathbf{\Omega} + \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}\right)^{-1}\left(\frac{1}{r_i}\mathbf{\Omega}\mathbf{u}_i + \frac{\mathbf{X}^T\mathbf{z}}{\sigma^2}\right) \\ \bar{\mathbf{\Omega}}_i &= \left(\frac{1}{r_i}\mathbf{\Omega} + \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}\right) \\ H &= \left(\frac{1}{r_i}\mathbf{u}_i^T\mathbf{\Omega}\mathbf{u}_i + \frac{\mathbf{z}^T\mathbf{z}}{\sigma^2}\right) - \left(\frac{1}{r_i}\mathbf{\Omega}\mathbf{u}_i + \frac{\mathbf{X}^T\mathbf{z}}{\sigma^2}\right)^T\left(\frac{1}{r_i}\mathbf{\Omega} + \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}\right)^{-1}\left(\frac{1}{r_i}\mathbf{\Omega}\mathbf{u}_i + \frac{\mathbf{X}^T\mathbf{z}}{\sigma^2}\right) \end{aligned}$$

Now let

$$\mathbf{y} = \sqrt{\bar{\mathbf{\Omega}}_i}\ (\mathbf{a} - \mathbf{v}_i),$$

so that the integral becomes

$$\begin{aligned} p(\omega_i|\mathbf{z}, \mathbf{X}) &\propto \frac{\pi_i\ e^{-H/2}}{(2\pi)^{N/2}\left|r_i\mathbf{\Omega}^{-1}\right|^{1/2}\left|\bar{\mathbf{\Omega}}_i\right|^{1/2}} \int \cdots \int_{\mathbb{R}^N} e^{-\frac{1}{2}\mathbf{y}^T\mathbf{y}}\ \mathrm{d}\mathbf{y}, \\ &= \frac{\pi_i\ e^{-H/2}}{\left|r_i\mathbf{\Omega}^{-1}\right|^{1/2}\left|\frac{1}{r_i}\mathbf{\Omega} + \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}\right|^{1/2}} \\ &= \frac{\pi_i\ e^{-H/2}}{\left|\mathbf{I}_N + \frac{r_i}{\sigma^2}\mathbf{\Omega}^{-1}\mathbf{X}^T\mathbf{X}\right|^{1/2}}. \end{aligned}$$

This can be simplified further by noting that in the exponent H, we can write

$$\begin{aligned} \mathbf{u}_i^T\mathbf{\Omega}\mathbf{u}_i &= \left[\mathbf{\Omega}^{-1}\sum_{j=1}^{C} q_{j,i}\mathbf{e}_j\right]^T \mathbf{\Omega}\left[\mathbf{\Omega}^{-1}\sum_{k=1}^{C} q_{k,i}\mathbf{e}_k\right], \\ &= \sum_{j=1}^{C} q_{j,i}\sum_{k=1}^{C}\mathbf{e}_j^T\mathbf{\Omega}^{-1}\mathbf{e}_k q_{k,i}, \\ &= \sum_{j=1}^{C} q_{j,i}\sum_{k=1}^{C} p_{j,k}q_{k,i}, \\ &= \sum_{j=1}^{C} q_{j,i}\delta_{i,j}, \\ &= q_{i,i}, \\ \mathbf{\Omega}\mathbf{u}_i &= \mathbf{\Omega}\mathbf{\Omega}^{-1}\sum_{j=1}^{C} q_{j,i}\mathbf{e}_j, \\ &= \sum_{j=1}^{C} q_{j,i}\mathbf{e}_j. \end{aligned}$$

Therefore,

$$p(\omega_i|\mathbf{z}, \mathbf{X}) \propto \frac{\pi_i\ e^{-H/2}}{\left|\mathbf{I}_N + \frac{r_i}{\sigma^2}\mathbf{\Omega}^{-1}\mathbf{X}^T\mathbf{X}\right|^{1/2}},$$

where

$$H = \frac{q_{i,i}}{r_i} + \frac{\mathbf{z}^T\mathbf{z}}{\sigma^2} - \left(\frac{1}{r_i}\sum_{j=1}^{C} q_{j,i}\mathbf{e}_j + \frac{\mathbf{X}^T\mathbf{z}}{\sigma^2}\right)^T \mathbf{M} \left(\frac{1}{r_i}\sum_{j=1}^{C} q_{j,i}\mathbf{e}_j + \frac{\mathbf{X}^T\mathbf{z}}{\sigma^2}\right),$$

$$\mathbf{M} = \left(\frac{1}{r_i}\mathbf{\Omega} + \frac{\mathbf{X}^T\mathbf{X}}{\sigma^2}\right)^{-1},$$

$$r_i = \frac{\kappa_i^2 - q_{i,i}}{N}.$$

To get the decision rule for implementation, we plug in the estimates for for $\mathbf{\Omega}$, κ_i^2, and σ^2, given in (5.56) - (5.58). It then simplifies to

$$p(\omega_i|\mathbf{z},\mathbf{X}) \propto \pi_i \ (\hat{\sigma}^2 + \hat{r}_i)^{-N/2} \ \exp\left\{-\frac{1}{2}\left[\frac{\hat{q}_{i,i}}{\hat{r}_i} + \frac{\mathbf{z}^T\mathbf{z}}{\hat{\sigma}^2}\right]\right\}$$
$$\times \ \exp\left\{\frac{\hat{\sigma}^2\hat{r}_i}{2(\hat{\sigma}^2+\hat{r}_i)}\left[\left(\sum_{j=1}^{C}\frac{\hat{q}_{j,i}}{\hat{r}_i}\mathbf{e}_j + \frac{\mathbf{X}^T\mathbf{z}}{\hat{\sigma}^2}\right)^T (\mathbf{X}^T\mathbf{X})^{-1}\left(\sum_{j=1}^{C}\frac{\hat{q}_{j,i}}{\hat{r}_i}\mathbf{e}_j + \frac{\mathbf{X}^T\mathbf{z}}{\hat{\sigma}^2}\right)\right]\right\}, \tag{3.32}$$

where $\hat{\mathbf{Q}} = (\mathcal{E}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathcal{E})^{-1}$, $\hat{r}_i = (\hat{\kappa}_i^2 - \hat{q}_{i,i})/N$, and $\hat{\kappa}_i^2$ is given in (5.57).

3.4 Hybrid Prior

3.4.1 Derivation

Another way of treating restrictions (1.10) and (1.11) is to require that the first one be satisfied with probability one, while the second is true only on the average, i.e.

$$P(\mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i) = 1,$$

$$\mathrm{E}_{\mathbf{a}|\omega_i}[\mathbf{a}^T\mathbf{\Omega}\mathbf{a}] = \kappa_i^2.$$

This represents a mixture of the Delta and Maximum Entropy interpretations, hence we will call it the Hybrid prior. The first condition is enforced by restricting the support of the prior to the intersection of the N hyperplanes (as in the Delta prior). Of all restricted priors, we choose the one with maximum entropy subject to $\mathrm{E}_{\mathbf{a}|\omega_i}[\mathbf{a}^T\mathbf{\Omega}\mathbf{a}] = \kappa_i^2$.

To restrict the domain of definition, we consider the set $\{\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^N,\ \mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i\}$. First note that the nullspace of $\mathcal{E}^T$ is spanned by the columns of the matrix

$$\mathbf{H} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{B}_C \end{pmatrix} \in \mathbb{R}^{N \times N-C}, \tag{4.33}$$

where

$$\mathbf{B}_i = \begin{pmatrix} -\mathbf{I}_{N_i-1} \\ \mathbf{1}^T \end{pmatrix} \in \mathbb{R}^{N_i \times N_i - 1}.$$

(Note that the columns of $\mathcal{E}$ and $\mathbf{H}$ form a basis for $\mathbb{R}^N$, and that $\mathcal{E}^T\mathbf{H} = \mathbf{0}$.) Next we form a particular solution to the equation $\mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i$, which is given by $\mathbf{a} = \mathbf{e}_i/N_i$. We therefore conclude that

$$\left\{\mathbf{a} \mid \mathbf{a} \in \mathbb{R}^N,\ \mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i\right\} = \left\{\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y} \mid \mathbf{y} \in \mathbb{R}^{N-C}\right\}. \tag{4.34}$$

As a result, the problem of specifying a prior on $\mathbf{a} \in \mathbb{R}^N$ is reduced to finding a prior on $\mathbf{y} \in \mathbb{R}^{N-C}$. The Hybrid prior $p(\mathbf{a}|\omega_i)$ is therefore the function $f_i(\mathbf{y})$ which maximizes

$$\mathrm{Ent}(f_i) = -\mathrm{E}[\ln f_i(\mathbf{y})] = -\int \cdots \int_{\mathbb{R}^{N-C}} f_i(\mathbf{y}) \ln f_i(\mathbf{y})\ \mathrm{d}\mathbf{y},$$

subject to the constraints

$$\begin{aligned} \kappa_i^2 &= \int \cdots \int_{\mathbb{R}^{N-C}} f_i(\mathbf{y}) \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right)^T \boldsymbol{\Omega} \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right)\ \mathrm{d}\mathbf{y}, \\ 1 &= \int \cdots \int_{\mathbb{R}^{N-C}} f_i(\mathbf{y})\ \mathrm{d}\mathbf{y}. \end{aligned} \tag{4.35}$$

We then seek to minimize the functional

$$\begin{aligned} F(f_i) = &\int \cdots \int_{\mathbb{R}^{N-C}} f_i(\mathbf{y}) \ln f_i(\mathbf{y})\ \mathrm{d}\mathbf{y} - \rho\left[\int \cdots \int_{\mathbb{R}^{N-C}} f_i(\mathbf{y})\ \mathrm{d}\mathbf{y} - 1\right] \\ &- \gamma\left[\int \cdots \int_{\mathbb{R}^{N-C}} f_i(\mathbf{y}) \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right)^T \boldsymbol{\Omega} \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right)\ \mathrm{d}\mathbf{a} - \kappa_i^2\right], \end{aligned}$$

where ρ and γ are the Lagrange multipliers. The Euler-Lagrange equation becomes

$$0 = \ln f_i(\mathbf{y}) + 1 - \rho - \gamma \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right)^T \boldsymbol{\Omega} \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right).$$

Then we have

$$f_i(\mathbf{y}) = C \exp\left\{ \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y} \right)^T \gamma \mathbf{\Omega} \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y} \right) \right\},$$

where C does not depend on $\mathbf{y}$. We can rewrite the exponent as

$$\begin{aligned} \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y} \right)^T \gamma \mathbf{\Omega} \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y} \right) &= \left[\mathbf{y} + (\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{\Omega} \mathbf{e}_i / N_i \right]^T \\ &\quad \times \gamma \mathbf{H}^T \mathbf{\Omega} \mathbf{H} \left[\mathbf{y} + (\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{\Omega} \mathbf{e}_i / N_i \right] \\ &\quad + D, \end{aligned}$$

where D is independent of $\mathbf{y}$. We then conclude that

$$\mathbf{y} | \omega_i \sim N\left(-(\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{\Omega} \mathbf{e}_i / N_i, \frac{(\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1}}{-2\gamma} \right). \tag{4.36}$$

To find γ, we use (4.35):

$$\begin{aligned} \kappa_i^2 &= \int \cdots \int_{\mathbb{R}^{N-C}} f_i(\mathbf{y}) \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y} \right)^T \mathbf{\Omega} \left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y} \right) \, \mathrm{d}\mathbf{y} \\ &= \mathrm{E}\left[\frac{1}{N_i^2} \mathbf{e}_i^T \mathbf{\Omega} \mathbf{e}_i + \frac{2}{N_i} \mathbf{e}_i^T \mathbf{\Omega} \mathbf{H} \mathbf{y} + \mathbf{y}^T \mathbf{H}^T \mathbf{\Omega} \mathbf{H} \mathbf{y} \right] \\ &= \frac{1}{N_i^2} \mathbf{e}_i^T \mathbf{\Omega} \mathbf{e}_i + \frac{2}{N_i} \mathbf{e}_i^T \mathbf{\Omega} \mathbf{H} \, \mathrm{E}\left[\mathbf{y} \right] + \mathrm{E}\left[\mathbf{y}^T \mathbf{H}^T \mathbf{\Omega} \mathbf{H} \mathbf{y} \right]. \end{aligned} \tag{4.37}$$

But from (4.36), we note that

$$\begin{aligned} \mathrm{E}\left[\mathbf{y} \right] &= -(\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{\Omega} \mathbf{e}_i / N_i, \\ \mathrm{E}\left[\mathbf{y}^T \mathbf{H}^T \mathbf{\Omega} \mathbf{H} \mathbf{y} \right] &= \mathrm{tr}\left[\frac{(\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1}}{-2\gamma} (\mathbf{H}^T \mathbf{\Omega} \mathbf{H}) \right] \\ &\quad + \frac{1}{N_i^2} \mathbf{e}_i^T \mathbf{\Omega} \mathbf{H} (\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-T} (\mathbf{H}^T \mathbf{\Omega} \mathbf{H}) (\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{\Omega} \mathbf{e}_i \\ &= \frac{N - C}{-2\gamma} + \frac{1}{N_i^2} \mathbf{e}_i^T \mathbf{\Omega} \mathbf{H} (\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-T} \mathbf{H}^T \mathbf{\Omega} \mathbf{e}_i. \end{aligned}$$

Plugging these into (4.37) and solving for γ yields

$$\frac{1}{-2\gamma} = \frac{\kappa_i^2 - \frac{1}{N_i^2} \left[\mathbf{e}_i^T \mathbf{\Omega} \mathbf{e}_i - \mathbf{e}_i^T \mathbf{\Omega} \mathbf{H} (\mathbf{H}^T \mathbf{\Omega} \mathbf{H})^{-1} \mathbf{H}^T \mathbf{\Omega} \mathbf{e}_i \right]}{N - C}.$$

With this value for γ, we write (4.36) as

$$\mathbf{y}|\omega_i \sim N(\boldsymbol{\mu}_{\mathbf{y}_i}, \boldsymbol{\Omega}_{\mathbf{y}_i}),$$

where we will use

$$\boldsymbol{\mu}_{\mathbf{y}_i} = -(\mathbf{H}^T\boldsymbol{\Omega}\mathbf{H})^{-1}\mathbf{H}^T\boldsymbol{\Omega}\mathbf{e}_i/N_i, \tag{4.38}$$

$$\alpha_{\mathbf{y}_i} = \frac{1}{N-C}\left[\kappa_i^2 - \frac{1}{N_i^2}\left(\mathbf{e}_i^T\boldsymbol{\Omega}\mathbf{e}_i - \mathbf{e}_i^T\boldsymbol{\Omega}\mathbf{H}(\mathbf{H}^T\boldsymbol{\Omega}\mathbf{H})^{-1}\mathbf{H}^T\boldsymbol{\Omega}\mathbf{e}_i\right)\right], \tag{4.39}$$

$$\boldsymbol{\Omega}_{\mathbf{y}_i} = (\mathbf{H}^T\boldsymbol{\Omega}\mathbf{H})^{-1}\alpha_{\mathbf{y}_i}. \tag{4.40}$$

The Hybrid prior (on the restricted vector $\mathbf{y} \in \mathbb{R}^{N-C}$) is then

$$p(\mathbf{y}|\omega_i) = \frac{1}{(2\pi)^{(N-C)/2}\left|\boldsymbol{\Omega}_{\mathbf{y}_i}\right|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}_i})^T\boldsymbol{\Omega}_{\mathbf{y}_i}^{-1}(\mathbf{y}-\boldsymbol{\mu}_{\mathbf{y}_i})\right\}. \tag{4.41}$$

Recall that this prior on $\mathbf{y}$ corresponds to the Hybrid prior on $\mathbf{a}|\omega_i$ according to set relation (4.34).

3.4.2 Decision Rule

As before, to compute the decision rule using the Hybrid prior, we evaluate (1.15) using $p(\mathbf{y}|\omega_i)$ in (4.41) and $p(\mathbf{z}|\mathbf{a},\mathbf{X})$ in (1.8). First, we write (1.8) in terms of the restricted vector $\mathbf{y}$:

$$\begin{aligned}
p(\mathbf{z}|\mathbf{a},\mathbf{X}) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{z}-\mathbf{X}\mathbf{a}\right)^T\left(\mathbf{z}-\mathbf{X}\mathbf{a}\right)\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{z}-\mathbf{X}\left(\frac{\mathbf{e}_i}{N_i}+\mathbf{H}\mathbf{y}\right)\right]^T\left[\mathbf{z}-\mathbf{X}\left(\frac{\mathbf{e}_i}{N_i}+\mathbf{H}\mathbf{y}\right)\right]\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{X}\mathbf{H}\mathbf{y}-\left(\mathbf{z}-\frac{1}{N_i}\mathbf{X}\mathbf{e}_i\right)\right]^T\left[\mathbf{X}\mathbf{H}\mathbf{y}-\left(\mathbf{z}-\frac{1}{N_i}\mathbf{X}\mathbf{e}_i\right)\right]\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\mathbf{y}^T\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}\mathbf{y} - 2\mathbf{y}^T\mathbf{H}^T\mathbf{X}^T(\mathbf{z}-\bar{\mathbf{x}}_i) + (\mathbf{z}-\bar{\mathbf{x}}_i)^T(\mathbf{z}-\bar{\mathbf{x}}_i)\right]\right\},
\end{aligned}$$

where

$$\bar{\mathbf{x}}_i = \frac{1}{N_i}\mathbf{X}\mathbf{e}_i \tag{4.42}$$

is the sample mean of the vectors from class ω_i. Next we plug in the above into (1.15) and integrate, giving

$$\begin{aligned} p(\omega_i|\mathbf{z},\mathbf{X}) &\propto \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{a}|\omega_i)\ p(\mathbf{z}|\mathbf{a},\mathbf{X})\ \mathrm{d}\mathbf{a} \\ &= \frac{\pi_i}{(2\pi\sigma^2)^{n/2}(2\pi)^{(N-C)/2}\left|\mathbf{\Omega}_{\mathbf{y}_i}\right|^{1/2}} \int \cdots \int_{\mathbb{R}^{N-C}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}^T\mathbf{A}\mathbf{y} - 2\mathbf{y}^T\mathbf{b} + c\right)\right\}\ \mathrm{d}\mathbf{y}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{A} &= \mathbf{\Omega}_{\mathbf{y}_i}^{-1} + \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}, \\ \mathbf{b} &= \mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i} + \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{X}^T(\mathbf{z} - \bar{\mathbf{x}}_i), \\ c &= \boldsymbol{\mu}_{\mathbf{y}_i}^T\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i} + \frac{1}{\sigma^2}(\mathbf{z} - \bar{\mathbf{x}}_i)^T(\mathbf{z} - \bar{\mathbf{x}}_i). \end{aligned}$$

The integral is evaluated from Lemma 3.2.3 and simplified, giving

$$\begin{aligned} p(\omega_i|\mathbf{z},\mathbf{X}) &\propto \frac{\pi_i}{(2\pi\sigma^2)^{n/2}(2\pi)^{(N-C)/2}\left|\mathbf{\Omega}_{\mathbf{y}_i}\right|^{1/2}} \left[\frac{(2\pi)^{N-C}\exp\left\{\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b} - c\right\}}{|\mathbf{A}|}\right]^{1/2} \\ &= \frac{\pi_i}{(2\pi\sigma^2)^{n/2}\left|\mathbf{I} + \frac{1}{\sigma^2}\mathbf{\Omega}_{\mathbf{y}_i}\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}\right|^{1/2}} \\ &\quad \times \exp\left\{-\frac{1}{2}\left[\boldsymbol{\mu}_{\mathbf{y}_i}^T\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i} + \frac{1}{\sigma^2}(\mathbf{z} - \bar{\mathbf{x}}_i)^T(\mathbf{z} - \bar{\mathbf{x}}_i)\right]\right\} \\ &\quad \times \exp\left\{\frac{1}{2}\ \boldsymbol{\mu}_{\mathbf{y}_i}^T\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\mathbf{A}^{-1}\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i}\right\} \\ &\quad \times \exp\left\{\frac{1}{\sigma^2}\boldsymbol{\mu}_{\mathbf{y}_i}^T\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\mathbf{A}^{-1}\mathbf{H}^T\mathbf{X}^T(\mathbf{z} - \bar{\mathbf{x}}_i)\right\} \\ &\quad \times \exp\left\{\frac{1}{2\sigma^4}(\mathbf{z} - \bar{\mathbf{x}}_i)^T\mathbf{X}\mathbf{H}\mathbf{A}^{-1}\mathbf{H}^T\mathbf{X}^T(\mathbf{z} - \bar{\mathbf{x}}_i)\right\}, \end{aligned} \tag{4.43}$$

where

$$\mathbf{A} = \mathbf{\Omega}_{\mathbf{y}_i}^{-1} + \frac{1}{\sigma^2}\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H}.$$

Equation (4.43) actually reduces to a much simpler form when we write it in terms of the original parameters. But rather than deriving the final form here, we will consider a more generalized setting in the next section.

3.4.3 Generalization

Recall that we have modeled the observed vector $\mathbf{z}$ as

$$\mathbf{z} = \mathbf{X}\mathbf{a} + \boldsymbol{\delta},$$

where $\mathbf{X}\mathbf{a}$ represents the component of $\mathbf{z}$ in $L(\mathbf{X})$, and $\boldsymbol{\delta}$ was treated as independent Gaussian noise. We will now consider alternate models for $\boldsymbol{\delta}$ to more accurately model the component of $\mathbf{z}$ in $L(\mathbf{X})^{\perp}$, while still using the Hybrid prior for $\mathbf{a}|\omega_i$. Specifically, we will replace the assumption $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ from (1.7) with the more general form

$$\boldsymbol{\delta} \sim N(\mathbf{0}, \boldsymbol{\Lambda}), \tag{4.44}$$

where $\boldsymbol{\Lambda}$ will allow us to impose some degree of orthogonality between $\mathbf{X}\mathbf{a}$ and $\boldsymbol{\delta}$. Hence the Hybrid prior on $\mathbf{a}$ will model the component of $\mathbf{z}$ in $L(\mathbf{X})$, and $\boldsymbol{\delta}$ will model the component of $\mathbf{z}$ in $L(\mathbf{X})^{\perp}$.

Perhaps the most straightforward approach is to require that $\boldsymbol{\delta} \perp \mathbf{X}$. While this ensures that the two components of $\mathbf{z}$ will be in orthogonal subspaces, the resulting decision rule will not give true probabilities. However, this approach will give us insight into a more appropriate choice for $\boldsymbol{\Lambda}$. Therefore, let us write $\boldsymbol{\Lambda} = \mathbf{U}\mathbf{U}^T$ so that $\boldsymbol{\delta} \sim N(\mathbf{0}, \mathbf{U}\mathbf{U}^T)$. We wish to find the form for $\mathbf{U}$ so $\boldsymbol{\delta} \perp L(\mathbf{X})$. Next we note that, since $\mathbf{X} \in \mathbb{R}^{n\times N}$, there exist an orthogonal matrix $\mathbf{O}_{\mathbf{X}} \in \mathbb{R}^{n\times n}$ and a matrix $\mathbf{B}_{\mathbf{X}} \in \mathbb{R}^{n\times N}$ of the form

$$\mathbf{B}_{\mathbf{X}} = \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{B}_0 \in \mathbb{R}^{N\times N}$, such that $\mathbf{X} = \mathbf{O}_\mathbf{X}\mathbf{B}_\mathbf{X}$. With this form for $\mathbf{X}$, we can easily establish the following:

Lemma 3.4.1. *If* $\boldsymbol{\delta} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$, *with*

$$\boldsymbol{\Lambda} = \mathbf{O}_\mathbf{X} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{L} \end{pmatrix} \mathbf{O}_\mathbf{X}^T, \tag{4.45}$$

for any (positive definite) $\mathbf{L} \in \mathbb{R}^{(n-N)\times(n-N)}$, *then* $\boldsymbol{\delta} \perp L(\mathbf{X})$.

Proof. Write $\mathbf{L} = \sqrt{\mathbf{L}}^{\,T}\sqrt{\mathbf{L}}$. Then for $\boldsymbol{\xi} \sim N(\mathbf{0}, \mathbf{I})$, we have

$$\boldsymbol{\delta} = \mathbf{O}_\mathbf{X} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt{\mathbf{L}}^{\,T} \end{pmatrix} \boldsymbol{\xi}.$$

Then

$$\begin{aligned} \boldsymbol{\delta}^T\mathbf{X} &= \boldsymbol{\xi}^T \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt{\mathbf{L}} \end{pmatrix} \mathbf{O}_\mathbf{X}^T\mathbf{O}_\mathbf{X}\mathbf{B}_\mathbf{X} \\ &= \boldsymbol{\xi}^T \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sqrt{\mathbf{L}} \end{pmatrix} \begin{pmatrix} \mathbf{B}_0 \\ \mathbf{0} \end{pmatrix} \\ &= \mathbf{0}. \end{aligned}$$

□

Let us now assume that $\boldsymbol{\Lambda}$ is of the form (4.45), so that $\boldsymbol{\delta} \perp \mathbf{X}$. But since $\boldsymbol{\delta} = \mathbf{z} - \mathbf{X}\mathbf{a}$, then we have $(\mathbf{z} - \mathbf{X}\mathbf{a})^T\mathbf{X} = \mathbf{0}$, so that $\mathbf{a}$ has a unique solution given by

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}.$$

This implies that

$$p(\mathbf{z}|\mathbf{a}, \mathbf{X}) = \delta\left(\mathbf{a} - (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}\right).$$

Combining this with the Hybrid prior $p_{\mathbf{a}|\omega_i}$, and plugging into (1.15), we obtain

$$p(\omega_i|\mathbf{z}, \mathbf{X}) \propto \pi_i \; p_{\mathbf{a}|\omega_i}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}\right),$$

Now recall that in the Hybrid prior on class ω_i, we require that $\mathcal{E}^T\mathbf{a} = \boldsymbol{\nu}_i$ with probability 1. As a result, $p_{\mathbf{a}|\omega_i}\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}\right) = 0$ unless the observed vector $\mathbf{z}$ satisfies $\mathcal{E}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z} = \boldsymbol{\nu}_i$, which will – in general – not be true for any i.

Hence choosing $\boldsymbol{\Lambda}$ of the form (4.45) is too extreme, since it will not give posteriors for each class. Therefore, to give meaningful posteriors, we will need a less restrictive choice. One option is to set $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_0 + \boldsymbol{\Lambda}_1$, where $\boldsymbol{\Lambda}_0$ is of the form (4.45), and $\boldsymbol{\Lambda}_1$ is chosen to avoid a restricted distribution, i.e. allow $\boldsymbol{\Lambda}$ to be invertible. To do this, we will set

$$\boldsymbol{\Lambda} = \mathbf{O}_{\mathbf{X}} \begin{pmatrix} d_1\mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & d_2\mathbf{I}_{n-N} \end{pmatrix} \mathbf{O}_{\mathbf{X}}^T. \tag{4.46}$$

Note that setting $d_1 = d_2 = \sigma^2$ results in the original Hybrid posteriors, and setting $d_1 = 0$ gives the degenerate case described above.

In this case, we can follow the derivation as in the original Hybrid posterior, where now (4.44) implies that $p(\mathbf{z}|\mathbf{a}, \mathbf{X}) = p(\mathbf{z}|\mathbf{a}, \mathbf{X}, \boldsymbol{\Lambda}) \sim N(\mathbf{X}\mathbf{a}, \boldsymbol{\Lambda})$. To derive the decision rule with the Hybrid prior on $\mathbf{a}|\omega_i$, we first rewrite $p(\mathbf{z}|\mathbf{a}, \mathbf{X}, \boldsymbol{\Lambda})$ in terms of the restricted vector $\mathbf{y}$ (see (4.34) and Section 3.4.2)

$$\begin{aligned}
p(\mathbf{z}|\mathbf{a}, \mathbf{X}, \boldsymbol{\Lambda}) &= \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Lambda}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{z} - \mathbf{X}\mathbf{a}\right)^T \boldsymbol{\Lambda}^{-1}\left(\mathbf{z} - \mathbf{X}\mathbf{a}\right)\right\} \\
&= \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Lambda}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left[\mathbf{z} - \mathbf{X}\left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right)\right]^T \boldsymbol{\Lambda}^{-1}\left[\mathbf{z} - \mathbf{X}\left(\frac{\mathbf{e}_i}{N_i} + \mathbf{H}\mathbf{y}\right)\right]\right\} \\
&= \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Lambda}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left[\mathbf{X}\mathbf{H}\mathbf{y} - \left(\mathbf{z} - \frac{1}{N_i}\mathbf{X}\mathbf{e}_i\right)\right]^T \boldsymbol{\Lambda}^{-1}\left[\mathbf{X}\mathbf{H}\mathbf{y} - \left(\mathbf{z} - \frac{1}{N_i}\mathbf{X}\mathbf{e}_i\right)\right]\right\} \\
&= \frac{1}{(2\pi)^{n/2}\left|\boldsymbol{\Lambda}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left[\mathbf{y}^T\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X}\mathbf{H}\mathbf{y} + (\mathbf{z} - \bar{\mathbf{x}}_i)^T\boldsymbol{\Lambda}^{-1}(\mathbf{z} - \bar{\mathbf{x}}_i)\right.\right. \\
&\quad \left.\left. -2\mathbf{y}^T\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}(\mathbf{z} - \bar{\mathbf{x}}_i)\right]\right\},
\end{aligned}$$

where $\bar{\mathbf{x}}_i$ was defined in (4.42). The decision rule is obtained by combining this with the

Hybrid prior $p(\mathbf{y}|\omega_i)$ in (4.41), giving

$$\begin{aligned} p(\omega_i|\mathbf{z},\mathbf{X}) &\propto \pi_i \int \cdots \int_{\mathbb{R}^N} p(\mathbf{a}|\omega_i)\ p(\mathbf{z}|\mathbf{a},\mathbf{X},\mathbf{\Lambda})\ \mathrm{d}\mathbf{a} \\ &= C_0 \int \cdots \int_{\mathbb{R}^{N-C}} \exp\left\{-\frac{1}{2}\left(\mathbf{y}^T\mathbf{A}_0\mathbf{y} - 2\mathbf{y}^T\mathbf{b}_0 + c_0\right)\right\}\ \mathrm{d}\mathbf{y}, \end{aligned}$$

where

$$\begin{aligned} C_0 &= \frac{\pi_i}{(2\pi)^{(n+N-C)/2}\left|\mathbf{\Lambda}\right|^{1/2}\left|\mathbf{\Omega}_{\mathbf{y}_i}\right|^{1/2}}, \\ \mathbf{A}_0 &= \mathbf{\Omega}_{\mathbf{y}_i}^{-1} + \mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}\mathbf{X}\mathbf{H}, \\ \mathbf{b}_0 &= \mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i} + \mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i), \\ c_0 &= \boldsymbol{\mu}_{\mathbf{y}_i}^T\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i} + (\mathbf{z}-\bar{\mathbf{x}}_i)^T\mathbf{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i). \end{aligned}$$

Using Lemma 3.2.3, we obtain

$$\begin{aligned} p(\omega_i|\mathbf{z},\mathbf{X}) &\propto C_0 \left[\frac{(2\pi)^N \exp\left\{\mathbf{b}_0^T\mathbf{A}_0^{-1}\mathbf{b}_0 - c_0\right\}}{|\mathbf{A}_0|}\right]^{1/2} \\ &\propto \frac{\pi_i\ \ e^{-(c_0-\mathbf{b}_0^T\mathbf{A}_0^{-1}\mathbf{b}_0)/2}}{\left|\mathbf{I}_{N-C}+\mathbf{\Omega}_{\mathbf{y}_i}\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}\mathbf{X}\mathbf{H}\right|^{1/2}}, \end{aligned}$$

where in the last line we have substituted in for C_0 and $|\mathbf{A}_0|$, and we dropped terms that are independent of i. Plugging in for $\mathbf{A}_0$, $\mathbf{b}_0$, and c_0, we obtain

$$\begin{aligned} p(\omega_i|\mathbf{z},\mathbf{X}) \propto\ & \frac{\pi_i}{\left|\mathbf{I}_{N-C}+\mathbf{\Omega}_{\mathbf{y}_i}\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}\mathbf{X}\mathbf{H}\right|^{1/2}}\ \exp\left\{\boldsymbol{\mu}_{\mathbf{y}_i}^T\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i} + (\mathbf{z}-\bar{\mathbf{x}}_i)^T\mathbf{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}^{-\frac{1}{2}} \\ & \times \exp\left\{\left(\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i}\right)^T\left(\mathbf{\Omega}_{\mathbf{y}_i}^{-1}+\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}\mathbf{X}\mathbf{H}\right)^{-1}\left(\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i}\right)\right]^{\frac{1}{2}} \\ & \times \exp\left\{\left(\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)^T\left(\mathbf{\Omega}_{\mathbf{y}_i}^{-1}+\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}\mathbf{X}\mathbf{H}\right)^{-1}\left(\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)\right\}^{\frac{1}{2}} \\ & \times \exp\left\{\left(\mathbf{\Omega}_{\mathbf{y}_i}^{-1}\boldsymbol{\mu}_{\mathbf{y}_i}\right)^T\left(\mathbf{\Omega}_{\mathbf{y}_i}^{-1}+\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}\mathbf{X}\mathbf{H}\right)^{-1}\left(\mathbf{H}^T\mathbf{X}^T\mathbf{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)\right\}. \end{aligned}$$

Next we substitute in for $\boldsymbol{\mu}_{\mathbf{y}_i}$ and $\boldsymbol{\Omega}_{\mathbf{y}_i}$, using (4.38) - (4.40), giving

$$
\begin{aligned}
p(\omega_i|\mathbf{z},\mathbf{X}) \propto & \frac{\pi_i}{|\mathbf{I}_{N-C}+\alpha_{\mathbf{y}_i}(\mathbf{H}^T\boldsymbol{\Omega}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X}\mathbf{H}|^{1/2}} \\
& \times \exp\left\{\frac{1}{\alpha_{\mathbf{y}_i}N_i^2}\,\mathbf{e}_i^T\boldsymbol{\Omega}\mathbf{H}(\mathbf{H}^T\boldsymbol{\Omega}\mathbf{H})^{-1}\mathbf{H}^T\boldsymbol{\Omega}\mathbf{e}_i+(\mathbf{z}-\bar{\mathbf{x}}_i)^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}^{-\frac{1}{2}} \\
& \times \exp\left\{\frac{1}{\alpha_{\mathbf{y}_i}^2N_i^2}\,\mathbf{e}_i^T\boldsymbol{\Omega}\mathbf{H}\left(\mathbf{H}^T\left(\alpha_{\mathbf{y}_i}^{-1}\boldsymbol{\Omega}+\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X}\right)\mathbf{H}\right)^{-1}\mathbf{H}^T\boldsymbol{\Omega}\mathbf{e}_i\right]^{\frac{1}{2}} \\
& \times \exp\left\{\left(\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)^T\left(\mathbf{H}^T\left(\alpha_{\mathbf{y}_i}^{-1}\boldsymbol{\Omega}+\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X}\right)\mathbf{H}\right)^{-1}\left(\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)\right\}^{\frac{1}{2}} \\
& \times \exp\left\{\frac{-1}{\alpha_{\mathbf{y}_i}N_i}\,\mathbf{e}_i^T\boldsymbol{\Omega}\mathbf{H}\left(\mathbf{H}^T\left(\alpha_{\mathbf{y}_i}^{-1}\boldsymbol{\Omega}+\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X}\right)\mathbf{H}\right)^{-1}\left(\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)\right\},
\end{aligned}
\tag{4.47}
$$

where $\alpha_{\mathbf{y}_i}$ is given by (4.39). Note that (4.47) determines the posteriors in terms of the true parameters $\boldsymbol{\Omega}$ and κ_i^2 (via $\alpha_{\mathbf{y}_i}$). In practice, these parameters will be estimated from the training samples, and as we shall see, this results in a much simpler expression.

Let us then plug in the estimates for $\boldsymbol{\Omega}$ and κ_i^2 from (5.56) and (5.57) into (4.47). Recalling that $\mathbf{X}\mathbf{e}_i/N_i = \bar{\mathbf{x}}_i$, it becomes

$$
\begin{aligned}
p(\omega_i|\mathbf{z},\mathbf{X}) \propto & \frac{\pi_i}{|\mathbf{I}_{N-C}+\hat{\alpha}_{\mathbf{y}_i}(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X}\mathbf{H}|^{1/2}} \\
& \times \exp\left\{\frac{1}{\hat{\alpha}_{\mathbf{y}_i}}\,\bar{\mathbf{x}}_i^T\mathbf{X}\mathbf{H}(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{X}^T\bar{\mathbf{x}}_i+(\mathbf{z}-\bar{\mathbf{x}}_i)^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}^{-\frac{1}{2}} \\
& \times \exp\left\{\frac{1}{\hat{\alpha}_{\mathbf{y}_i}^2}\,\bar{\mathbf{x}}_i^T\mathbf{X}\mathbf{H}\left(\mathbf{H}^T\mathbf{X}^T\left(\hat{\alpha}_{\mathbf{y}_i}^{-1}+\boldsymbol{\Lambda}^{-1}\right)\mathbf{X}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{X}^T\bar{\mathbf{x}}_i^T\right]^{\frac{1}{2}} \\
& \times \exp\left\{\left(\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)^T\left(\mathbf{H}^T\mathbf{X}^T\left(\hat{\alpha}_{\mathbf{y}_i}^{-1}+\boldsymbol{\Lambda}^{-1}\right)\mathbf{X}\mathbf{H}\right)^{-1}\left(\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)\right\}^{\frac{1}{2}} \\
& \times \exp\left\{\frac{-1}{\hat{\alpha}_{\mathbf{y}_i}}\,\bar{\mathbf{x}}_i^T\mathbf{X}\mathbf{H}\left(\mathbf{H}^T\mathbf{X}^T\left(\hat{\alpha}_{\mathbf{y}_i}^{-1}+\boldsymbol{\Lambda}^{-1}\right)\mathbf{X}\mathbf{H}\right)^{-1}\left(\mathbf{H}^T\mathbf{X}^T\boldsymbol{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{x}}_i)\right)\right\},
\end{aligned}
$$

where $\hat{\alpha}_{\mathbf{y}_i}$ is obtained from (4.39), giving

$$\begin{aligned}\hat{\alpha}_{\mathbf{y}_i} &= \frac{1}{N-C}\left[\hat{\kappa}_i^2 - \frac{1}{N_i^2}\left(\mathbf{e}_i^T\hat{\mathbf{\Omega}}\mathbf{e}_i - \mathbf{e}_i^T\hat{\mathbf{\Omega}}\mathbf{H}(\mathbf{H}^T\hat{\mathbf{\Omega}}\mathbf{H})^{-1}\mathbf{H}^T\hat{\mathbf{\Omega}}\mathbf{e}_i\right)\right] \\ &= \frac{1}{N-C}\left[\hat{\sigma}_i^2 + \|\bar{\mathbf{x}}_i\|^2 - \|\bar{\mathbf{x}}_i\|^2 + \bar{\mathbf{x}}_i^T\mathbf{X}\mathbf{H}(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{X}^T\bar{\mathbf{x}}_i\right] \\ &= \frac{1}{N-C}\left[\hat{\sigma}_i^2 + \bar{\mathbf{x}}_i^T\mathbf{X}\mathbf{H}(\mathbf{H}^T\mathbf{X}^T\mathbf{X}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{X}^T\bar{\mathbf{x}}_i\right].\end{aligned}$$

Now we define the matrix

$$\mathbf{Y} = \mathbf{X}\mathbf{H}, \tag{4.48}$$

so that we can write is more compactly as

$$\begin{aligned}p(\omega_i|\mathbf{z},\mathbf{X}) \propto & \frac{\pi_i}{|\mathbf{I}_{N-C} + \hat{\alpha}_{\mathbf{y}_i}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{\Lambda}^{-1}\mathbf{Y}|^{1/2}} \\ & \times \exp\left\{(\mathbf{z}-\bar{\mathbf{x}}_i)^T\left(\mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{Y}\left[\mathbf{Y}^T\left(\hat{\alpha}_{\mathbf{y}_i}^{-1}\mathbf{I}_n + \mathbf{\Lambda}^{-1}\right)\mathbf{Y}\right]^{-1}\mathbf{Y}^T\mathbf{\Lambda}^{-1}\right)(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}^{-1/2} \\ & \times \exp\left\{\bar{\mathbf{x}}_i^T\left(\hat{\alpha}_{\mathbf{y}_i}^{-1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T - \hat{\alpha}_{\mathbf{y}_i}^{-2}\mathbf{Y}\left[\mathbf{Y}^T\left(\hat{\alpha}_{\mathbf{y}_i}^{-1}\mathbf{I}_n + \mathbf{\Lambda}^{-1}\right)\mathbf{Y}\right]^{-1}\mathbf{Y}^T\right)\bar{\mathbf{x}}_i\right\}^{-1/2} \\ & \times \exp\left\{\bar{\mathbf{x}}_i^T\left(2\hat{\alpha}_{\mathbf{y}_i}^{-1}\mathbf{Y}\left[\mathbf{Y}^T\left(\hat{\alpha}_{\mathbf{y}_i}^{-1}\mathbf{I}_n + \mathbf{\Lambda}^{-1}\right)\mathbf{Y}\right]^{-1}\mathbf{Y}^T\mathbf{\Lambda}^{-1}\right)(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}^{-1/2},\end{aligned} \tag{4.49}$$

where

$$\hat{\alpha}_{\mathbf{y}_i} = \frac{1}{N-C}\left[\hat{\sigma}_i^2 + \bar{\mathbf{x}}_i^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\bar{\mathbf{x}}_i\right].$$

Next, we will use our choice for $\mathbf{\Lambda}$ given in (4.46). Note that since $\mathbf{Y} = \mathbf{X}\mathbf{H}$, and $\mathbf{X} = \mathbf{O}_\mathbf{X}\mathbf{B}_\mathbf{X}$, we have the useful result

$$\begin{aligned}\mathbf{\Lambda}^{-1}\mathbf{Y} &= \mathbf{O}_\mathbf{X}\begin{pmatrix} d_1^{-1}\mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & d_2^{-1}\mathbf{I}_{n-N}\end{pmatrix}\mathbf{O}_\mathbf{X}^T\mathbf{O}_\mathbf{X}\mathbf{B}_\mathbf{X}\mathbf{H} \\ &= d_1^{-1}\mathbf{O}_\mathbf{X}\mathbf{B}_\mathbf{X}\mathbf{H} \\ &= d_1^{-1}\mathbf{X}\mathbf{H} \\ &= d_1^{-1}\mathbf{Y}.\end{aligned}$$

As a result, (4.49) becomes

$$p(\omega_i|\mathbf{z},\mathbf{X}) \propto \pi_i(\hat{\alpha}_{\mathbf{y}_i}+d_1)^{-\frac{N-C}{2}} \exp\left\{(\mathbf{z}-\bar{\mathbf{x}}_i)^T\left(\mathbf{\Lambda}^{-1}-\frac{\hat{\alpha}_{\mathbf{y}_i}/d_1}{\hat{\alpha}_{\mathbf{y}_i}+d_1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\right)(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}^{-1/2}$$
$$\times \exp\left\{\bar{\mathbf{x}}_i^T\left(\frac{1}{\hat{\alpha}_{\mathbf{y}_i}+d_1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\right)\bar{\mathbf{x}}_i\right\}^{-1/2}$$
$$\times \exp\left\{\bar{\mathbf{x}}_i^T\left(\frac{2}{\hat{\alpha}_{\mathbf{y}_i}+d_1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\right)(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}^{-1/2}. \tag{4.50}$$

Note that the exponent is a quadratic in $(\mathbf{z}-\bar{\mathbf{x}}_i)$ and $\bar{\mathbf{x}}_i$. We therefore seek a convenient factorization in hopes of gaining insight into behavior of the classifier. This will require the following Lemma:

Lemma 3.4.2. *For any symmetric* $\mathbf{A},\mathbf{B},\mathbf{C}\in\mathbb{R}^{n\times n}$*, we have*

$$(\mathbf{z}-\bar{\mathbf{x}}_i)^T\mathbf{A}(\mathbf{z}-\bar{\mathbf{x}}_i)+\bar{\mathbf{x}}_i^T\mathbf{B}\bar{\mathbf{x}}_i+\bar{\mathbf{x}}_i^T\mathbf{C}(\mathbf{z}-\bar{\mathbf{x}}_i)=(\mathbf{C}'\mathbf{z}-\bar{\mathbf{x}}_i)^T\mathbf{A}'(\mathbf{C}'\mathbf{z}-\bar{\mathbf{x}}_i)+\mathbf{z}^T\mathbf{B}'\mathbf{z},$$

if and only if

$$\mathbf{A}'=\mathbf{A}+\mathbf{B}-\mathbf{C}$$
$$\mathbf{B}'=\mathbf{A}-(\mathbf{A}-\mathbf{C}/2)(\mathbf{A}+\mathbf{B}-\mathbf{C})^{-1}(\mathbf{A}-\mathbf{C}/2)$$
$$\mathbf{C}'=(\mathbf{A}+\mathbf{B}-\mathbf{C})^{-1}(\mathbf{A}-\mathbf{C}/2).$$

Comparing Lemma 3.4.2 with the exponent in (4.50), and noting in this case that $2\mathbf{B}=\mathbf{C}$, we can easily obtain

$$\mathbf{A}'=\mathbf{\Lambda}^{-1}-d_1^{-1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$$
$$\mathbf{B}'=\frac{1}{\hat{\alpha}_{\mathbf{y}_i}+d_1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$$
$$\mathbf{C}'=\mathbf{I}.$$

We can then rewrite (4.50) as

$$p(\omega_i|\mathbf{z},\mathbf{X}) \propto \pi_i(\hat{\alpha}_{\mathbf{y}_i} + d_1)^{-\frac{N-C}{2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\bar{\mathbf{x}}_i)^T\left(\mathbf{\Lambda}^{-1} - d_1^{-1}\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\right)(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}$$
$$\times \exp\left\{-\frac{1}{2(\hat{\alpha}_{\mathbf{y}_i}+d_1)}\,\mathbf{z}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{z}\right\}.$$

Now, observe that the matrix $\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$ projects onto the column space of $\mathbf{Y}$, i.e. $\text{proj}_{L(\mathbf{Y})}\mathbf{z} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{z}$. Hence we will define the matrix

$$\mathbf{P}_{\mathbf{Y}} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T.$$

This allows us to write the generalized Hybrid posteriors as

$$p(\omega_i|\mathbf{z},\mathbf{X}) \propto \pi_i(\hat{\alpha}_{\mathbf{y}_i} + d_1)^{-\frac{N-C}{2}} \exp\left\{-\frac{1}{2}(\mathbf{z}-\bar{\mathbf{x}}_i)^T\left(\mathbf{\Lambda}^{-1} - d_1^{-1}\mathbf{P}_{\mathbf{Y}}\right)(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}$$
$$\times \exp\left\{-\frac{1}{2(\hat{\alpha}_{\mathbf{y}_i}+d_1)}\,\mathbf{z}^T\mathbf{P}_{\mathbf{Y}}\mathbf{z}\right\}, \tag{4.51}$$

where now we can express $\hat{\alpha}_{\mathbf{y}_i}$ as

$$\hat{\alpha}_{\mathbf{y}_i} = \frac{1}{N-C}\left[\hat{\sigma}_i^2 + \|\mathbf{P}_{\mathbf{Y}}\bar{\mathbf{x}}_i\|^2\right].$$

The original Hybrid posteriors are obtained by setting $\mathbf{\Lambda} = \sigma^2\mathbf{I}$ in (4.51), which can be accomplished by setting $d_1 = d_2 = \sigma^2$ in (4.46). In this case, we obtain the original Hybrid posteriors, which can now be written

$$p(\omega_i|\mathbf{z},\mathbf{X}) \propto \pi_i(\hat{\alpha}_{\mathbf{y}_i} + \hat{\sigma}^2)^{-\frac{N-C}{2}} \exp\left\{-\frac{1}{2\hat{\sigma}^2}(\mathbf{z}-\bar{\mathbf{x}}_i)^T\mathbf{P}_{\mathbf{Y}^\perp}(\mathbf{z}-\bar{\mathbf{x}}_i)\right\}$$
$$\times \exp\left\{-\frac{1}{2(\hat{\alpha}_{\mathbf{y}_i}+\hat{\sigma}^2)}\,\mathbf{z}^T\mathbf{P}_{\mathbf{Y}}\mathbf{z}\right\},$$

where $\hat{\sigma}^2$ is given in (5.58), and $\mathbf{P}_{\mathbf{Y}^\perp} = \mathbf{I} - \mathbf{P}_{\mathbf{Y}}$ is the projection onto $L(\mathbf{Y})^\perp$. But since

projections are self-adjoint and idempotent, then

$$\begin{aligned}\mathbf{z}^T\mathbf{P_Y z} &=< \mathbf{z}, \mathbf{P_Y z} > \\ &=< \mathbf{z}, \mathbf{P_Y P_Y z} > \\ &=< \mathbf{P_Y z}, \mathbf{P_Y z} > \\ &= \|\mathbf{P_Y z}\|^2.\end{aligned}$$

Similarly, $(\mathbf{z} - \bar{\mathbf{x}}_i)^T\mathbf{P}_{\mathbf{Y}^\perp}(\mathbf{z} - \bar{\mathbf{x}}_i) = \|\mathbf{P}_{\mathbf{Y}^\perp}(\mathbf{z} - \bar{\mathbf{x}}_i)\|^2$. Hence, the original Hybrid posteriors have the convenient representation

$$p(\omega_i|\mathbf{z}, \mathbf{X}) \propto \pi_i(\hat{\alpha}_{\mathbf{y}_i} + \hat{\sigma}^2)^{-\frac{N-C}{2}} \ \exp\left\{-\frac{1}{2}\left(\frac{\|\mathbf{P}_{\mathbf{Y}^\perp}(\mathbf{z} - \bar{\mathbf{x}}_i)\|^2}{\hat{\sigma}^2} + \frac{\|\mathbf{P_Y z}\|^2}{\hat{\alpha}_{\mathbf{y}_i} + \hat{\sigma}^2}\right)\right\}. \tag{4.52}$$

3.5 Estimating Parameters

In this section, we consider how to estimate the parameters that were used in the decision rules. Recall that

$$\begin{aligned}\mathbf{X} &\sim N(\mathbf{M}\mathcal{E}^T, \mathbf{\Sigma} \otimes \mathbf{\Psi}), \\ \text{diag}(\mathbf{\Psi}) &= (\sigma_1^2, \ldots, \sigma_1^2, \sigma_2^2, \ldots, \sigma_2^2, \ldots, \sigma_C^2, \ldots, \sigma_C^2),\end{aligned}$$

where $\mathbf{X}$ is the matrix of training samples, $\mathbf{M}$ is the matrix of class-means, and $\mathcal{E}$ was defined in (1.3). To obtain the parameters σ_i^2, we first estimate $\mathbf{M}$ in the usual way by setting

$$\begin{aligned}\hat{\mathbf{M}} &= [\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots \bar{\mathbf{x}}_C] \\ &= \mathbf{X}\mathcal{E}\mathbf{D}_N^{-1},\end{aligned}$$

where $\bar{\mathbf{x}}_i$ is the sample mean from ω_i given in (4.42), and $\mathbf{D}_N = \text{diag}(N_1, N_2, \ldots, N_C)$. From the properties of matrix-normal distributions, we have

$$\begin{aligned}\text{E}(\mathbf{X}^T\mathbf{X}) &= \text{tr}(\boldsymbol{\Sigma})\boldsymbol{\Psi} + \text{E}(\mathbf{X})^T\text{E}(\mathbf{X}) \\ &= \text{tr}(\boldsymbol{\Sigma})\boldsymbol{\Psi} + (\mathbf{M}\mathcal{E}^T)^T(\mathbf{M}\mathcal{E}^T).\end{aligned}$$

Therefore,

$$\boldsymbol{\Psi} = \frac{1}{\text{tr}(\boldsymbol{\Sigma})}\left[\text{E}(\mathbf{X}^T\mathbf{X}) - \mathcal{E}\mathbf{M}^T\mathbf{M}\mathcal{E}^T\right].$$

But since (1.5) is invariant under the transformation $\boldsymbol{\Sigma} \to k\boldsymbol{\Sigma}, \boldsymbol{\Psi} \to k^{-1}\boldsymbol{\Psi}$, we can assume $\text{tr}(\boldsymbol{\Sigma}) = 1$. Then based on the single observed matrix $\mathbf{X}$, we estimate $\boldsymbol{\Psi}$ by

$$\begin{aligned}\hat{\boldsymbol{\Psi}} &= \mathbf{X}^T\mathbf{X} - \mathcal{E}\hat{\mathbf{M}}^T\hat{\mathbf{M}}\mathcal{E}^T \\ &= \mathbf{X}^T\mathbf{X} - \mathcal{E}\mathbf{D}_N^{-1}\mathcal{E}^T\mathbf{X}^T\mathbf{X}\mathcal{E}\mathbf{D}_N^{-1}\mathcal{E}^T \qquad (5.53)\end{aligned}$$

Next, consider the problem of estimating $\sigma_i^2,\ i = 1, \ldots, C$, given $\hat{\boldsymbol{\Psi}}$. Of course if $\hat{\boldsymbol{\Psi}}$ has appropriately sized blocks of constants along the diagonal, then we would choose $\hat{\sigma}_i^2$ to be these constants, and there would be no problem. However, this will generally not be the case because $\hat{\boldsymbol{\Psi}}$ in (5.53) will not have the appropriately structured diagonal. Note, though, that $\mathcal{E}\hat{\mathbf{M}}^T\hat{\mathbf{M}}\mathcal{E}^T$ is of the correct form, because

$$\text{diag}(\mathcal{E}\hat{\mathbf{M}}^T\hat{\mathbf{M}}\mathcal{E}^T) = (\|\bar{\mathbf{x}}_1\|^2, \ldots, \|\bar{\mathbf{x}}_1\|^2, \|\bar{\mathbf{x}}_2\|^2, \ldots, \|\bar{\mathbf{x}}_2\|^2, \ldots, \|\bar{\mathbf{x}}_C\|^2, \ldots, \|\bar{\mathbf{x}}_C\|^2).$$

Therefore, if we can force $\mathbf{X}^T\mathbf{X}$ to have a block-constant diagonal, then we will be able to obtain unambiguous estimates for the σ_i^2's.

To do this, let us preprocess all of the data (training and test) by normalizing each vector to have a norm of 1. Then we will have $\text{diag}(\mathbf{X}^T\mathbf{X}) = (1, 1, \ldots, 1)$ and $\hat{\boldsymbol{\Psi}}$ will have

the desired form without affecting any of the calculations. This yields

$$\hat{\sigma}_i^2 = 1 - \|\bar{\mathbf{x}}_i\|^2, \ i = 1, \ldots, C. \tag{5.54}$$

The drawback to this technique is that it does affect our assumption that $\mathbf{X} \sim N(\mathbf{M}\mathcal{E}^T, \mathbf{\Sigma} \otimes \mathbf{\Psi})$, because now the vectors in $\mathbf{X}$ would be projected onto the unit sphere, nullifying the normality assumption. Despite this, the numerical simulations often have better results for the normalized data, indicating that the improvement in the estimates for σ_i^2 counteract the deviation from normality. For simulations where the data is not normalized, it would be natural to choose $\hat{\sigma}_i^2$ to be the average of the appropriate diagonal block of $\hat{\mathbf{\Psi}}$, i.e.

$$\hat{\sigma}_i^2 = \text{diag}(\hat{\mathbf{\Psi}})^T \mathbf{e}_i / N_i, \ i = 1, \ldots, C. \tag{5.55}$$

Note that (5.55) reduces to (5.54) when the data is normalized.

The decision rules implemented in this paper use the parameters $\mathbf{\Omega}$ and κ_i^2, defined in (1.12) and (1.13). Their estimates are given by

$$\hat{\mathbf{\Omega}} = \mathbf{X}^T \mathbf{X}, \tag{5.56}$$

$$\hat{\kappa}_i^2 = \hat{\sigma}_i^2 + \|\bar{\mathbf{x}}_i\|^2. \tag{5.57}$$

It is worth noting that when the data is normalized, $\hat{\kappa}_i^2 = 1$.

When a new vector $\mathbf{z}$ is observed, we estimate the unknown parameter σ^2 in (1.8) with the MLE

$$\hat{\sigma}^2 = \|\mathbf{z} - \mathbf{X}\hat{\mathbf{a}}\|^2 / n, \tag{5.58}$$

where $\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{z} - \mathbf{X}\hat{\mathbf{a}}\|$.

3.6 Relation to Linear SVMs

Note that formula (1.6) implies that the vector $\mathbf{a}$ can be viewed as the coefficients of the projection of $\mathbf{z}$ onto $L(\mathbf{X}) = \text{span}\{\mathbf{x}_{i,j}\}$. Let us find these coefficients explicitly in the deterministic paradigm. If $\mathbf{u} = \text{proj}_{L(\mathbf{X})}\mathbf{z} = \mathbf{X}\mathbf{a}$, then $\mathbf{X}^T(\mathbf{z} - \mathbf{u}) = \mathbf{0}$, which leads to

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{z}. \tag{6.59}$$

Now consider the two-class problem and construct a classification rule based on a linear discriminant function

$$f(\mathbf{z}) = \mathbf{w}^T\mathbf{z}.$$

We classify a new vector $\mathbf{z}$ into class ω_1 whenever $f(\mathbf{z}) > 0$ (and ω_2 otherwise). For the training vectors, we want $\mathbf{w}^T\mathbf{x}_{1,j} = y_1$ and $\mathbf{w}^T\mathbf{x}_{2,j} = y_2$, which can be written

$$\mathbf{w}^T\mathbf{X} = \mathbf{y}^T, \qquad \mathbf{y}^T = (\underbrace{y_1, \ldots, y_1}_{N_1}, \underbrace{y_2, \ldots, y_2}_{N_2}). \tag{6.60}$$

We add the condition $\|\mathbf{w}\|^2 = \min$ to minimize the length of the vector $\mathbf{w}$ (and therefore maximize the margin). The optimization problem is therefore

$$\|\mathbf{w}\|^2 = \min,$$

$$\mathbf{X}^T\mathbf{w} = \mathbf{y}.$$

Solving this using Lagrange multipliers, we wish to minimize

$$L = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{2}\sum_{j=1}^{2}\lambda_{i,j}\left(\mathbf{x}_{i,j}^T\mathbf{w} - y_i\right). \tag{6.61}$$

If we compare this to equation (2.8) from Chapter 2, we see that it is identical to the linear SVM formulation with the additional requirement that all of the training samples must be

support vectors (i.e. they must all lie on the margin). And since we assume $n >> N$, the separating hyperplane still has $n - N$ degrees of freedom to maximize this margin.

If we write $\boldsymbol{\lambda}^T = (\lambda_{1,1}, \ldots, \lambda_{1,N_1}, \lambda_{2,1}, \ldots, \lambda_{2,N_2})$, then we can rewrite (6.61) as

$$L_a = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \boldsymbol{\lambda}^T \left(\mathbf{X}^T\mathbf{w} - \mathbf{b}\right).$$

Taking the derivative with respect to $\mathbf{w}$ we obtain

$$\mathbf{w} = \mathbf{X}\boldsymbol{\lambda}.$$

Combining this with (6.60) gives $\boldsymbol{\lambda}^T = \mathbf{y}^T(\mathbf{X}^T\mathbf{X})^{-1}$, so we obtain

$$\mathbf{w} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{y}$$

and

$$\begin{aligned} f(\mathbf{z}) &= \mathbf{w}^T\mathbf{z} \\ &= \mathbf{y}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{z} \\ &= \mathbf{y}^T\mathbf{a}, \end{aligned}$$

where $\mathbf{a}$ is the vector of coefficients of the projection of $\mathbf{z}$ onto $L(\mathbf{X})$ in (6.59). In SVMs, usually $y_1 = 1$ and $y_2 = -1$. Hence $f(\mathbf{z}) = \mathbf{e}_1^T\mathbf{a} - \mathbf{e}_2^T\mathbf{a}$, and $\mathbf{z}$ is classified into class ω_1 if $\mathbf{e}_1^T\mathbf{a} > \mathbf{e}_2^T\mathbf{a}$. Note that under our priors,

$$\mathbf{e}_1^T\mathbf{a} = 1, \qquad \mathbf{e}_2^T\mathbf{a} = 0,$$

for class ω_1, and

$$\mathbf{e}_1^T\mathbf{a} = 0, \qquad \mathbf{e}_2^T\mathbf{a} = 1,$$

for class ω_2. Hence, under the Delta and Hybrid priors, $f(\mathbf{z}) = 1$ for class ω_1, and $f(\mathbf{z}) = -1$ for class ω_2. This is consistent with the linear SVM classification method. The condition $\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a} = \kappa_i^2$ has the role of normalizing the data in our case.

3.7 Proofs

Proof of Lemma 3.2.2

Proof. Define the matrix $\mathbf{U}$ to be

$$\mathbf{U} = \begin{bmatrix} \mathbf{I}_{N-1} \\ -\mathbf{v}_l^T / v_N \end{bmatrix} \in \mathbb{R}^{N \times N-1}.$$

Therefore, we have

$$\mathbf{x}|_{x_N = \frac{1}{v_N}(c - x_1 v_1 - \ldots - x_{N-1} v_{N-1})} = \mathbf{U}\mathbf{x}_l + \frac{c}{v_N}\boldsymbol{\nu}_N,$$

so that (2.17) becomes

$$\begin{aligned} \mathbf{x}^T \boldsymbol{\Psi} \mathbf{x} + 2\mathbf{b}^T \mathbf{x} &= \left[\mathbf{U}\mathbf{x}_l + \frac{c}{v_N}\boldsymbol{\nu}_N\right]^T \boldsymbol{\Psi} \left[\mathbf{U}\mathbf{x}_l + \frac{c}{v_N}\boldsymbol{\nu}_N\right] + 2\mathbf{b}^T \left[\mathbf{U}\mathbf{x}_l + \frac{c}{v_N}\boldsymbol{\nu}_N\right] \\ &= \mathbf{x}_l^T \left[\mathbf{U}^T \boldsymbol{\Psi} \mathbf{U}\right] \mathbf{x}_l + 2\mathbf{x}_l^T \left[\frac{c}{v_N}\mathbf{U}^T \boldsymbol{\Psi}\boldsymbol{\nu}_N + \mathbf{U}^T \mathbf{b}\right] + \left[\frac{c^2}{v_N^2}\boldsymbol{\nu}_N^T \boldsymbol{\Psi}\boldsymbol{\nu}_N + 2\frac{c}{v_N}\mathbf{b}^T \boldsymbol{\nu}_N\right]. \end{aligned}$$

By comparing this to the right-hand side of (2.17), we have

$$\begin{aligned} \boldsymbol{\Phi} &= \mathbf{U}^T \boldsymbol{\Psi} \mathbf{U}, \\ \mathbf{w} &= \frac{c}{v_N}\mathbf{U}^T \boldsymbol{\Psi}\boldsymbol{\nu}_N + \mathbf{U}^T \mathbf{b}, \\ C &= \frac{c^2}{v_N^2}\boldsymbol{\nu}_N^T \boldsymbol{\Psi}\boldsymbol{\nu}_N + 2\frac{c}{v_N}\mathbf{b}^T \boldsymbol{\nu}_N. \end{aligned}$$

If follows by direct calculation that

$$\begin{aligned} \Phi_{i,j} &= \Psi_{i,j} - \frac{v_i \Psi_{N,j} + v_j \Psi_{N,i}}{v_N} + \frac{v_i v_j}{v_N^2}\Psi_{N,N},\ i,j = 1,\ldots,N-1, \\ w_i &= \frac{c}{v_N}\left(\Psi_{N,i} - \frac{v_i \Psi_{N,N}}{v_N}\right) + b_i - \frac{b_N}{v_N}v_i,\ i = 1,\ldots,N-1, \\ C &= \frac{\Psi_{N,N}}{v_N^2}c^2 + \frac{2b_N}{v_N}c. \end{aligned}$$

Now let us assume that $\boldsymbol{\Psi}$ is positive definite. Let $\mathbf{r} = (r_1,\ldots,r_{N-1})^T \in \mathbb{R}^{N-1}$, $\mathbf{r} \neq \mathbf{0}$. Then $\mathbf{r}^T \boldsymbol{\Phi} \mathbf{r} = \mathbf{s}^T \boldsymbol{\Psi} \mathbf{s}$, where

$$\mathbf{s} = (r_1, r_2, \ldots, r_{N-1}, -r_1\frac{v_1}{v_N} - \ldots - r_{N-1}\frac{v_{N-1}}{v_N})^T \in \mathbb{R}^N.$$

But then $\mathbf{s} \neq \mathbf{0}$, so $\mathbf{s}^T \mathbf{\Phi} \mathbf{s} > 0$. Thus $\mathbf{r}^T \mathbf{\Phi} \mathbf{r} > 0$ for all $\mathbf{r} \neq \mathbf{0}$.

Now consider the particular case where $\mathbf{\Psi} = \mathbf{I}_N$, $\mathbf{b} = \mathbf{0}$, and $v_N = 1$. In this case, (2.18) - (2.21) are easily checked. To see (2.22), note that $\mathbf{\Phi} - \mathbf{I}_{N-1} = \mathbf{v}_s \mathbf{v}_s^T$, which has rank 1. Thus 1 is an eigenvalue of $\mathbf{\Phi}$ with multiplicity $N-2$. Furthermore, it is easy to check that $\|\mathbf{v}\|^2$ is the remaining eigenvalue. Thus $|\mathbf{\Phi}| = \|\mathbf{v}\|^2$.

□

Proof of Lemma 3.2.3

Proof.

$$\begin{aligned}\int \cdots \int_{\mathbb{R}^N} e^{-\frac{1}{2}\left(\mathbf{x}^T\mathbf{A}\mathbf{x}+2\mathbf{b}^T\mathbf{x}+c\right)}\, \mathrm{d}\mathbf{x} &= e^{-\frac{1}{2}\left(c-\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}\right)} \int \cdots \int_{\mathbb{R}^N} e^{-\frac{1}{2}\left[\left(\mathbf{x}+\mathbf{A}^{-1}\mathbf{b}\right)^T\mathbf{A}\left(\mathbf{x}+\mathbf{A}^{-1}\mathbf{b}\right)\right]}\, \mathrm{d}\mathbf{x} \\ &= \left[\frac{(2\pi)^N \exp\left\{\mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}-c\right\}}{|\mathbf{A}|}\right]^{1/2}\end{aligned}$$

□

Proof of Lemma 3.2.4

Proof. To prove this Lemma, we first establish that for $\mathbf{x}, \mathbf{v} \in \mathbb{R}^N$, $\mathbf{v} \neq \mathbf{0}$, $c \in \mathbb{R}$,

$$\int \cdots \int_{\mathbb{R}^N} e^{-\mathbf{x}^T\mathbf{x}}\, \delta(\mathbf{v}^T\mathbf{x}-c)\, \mathrm{d}\mathbf{x} = \frac{\sqrt{\pi}^{\,N-1}}{\|\mathbf{v}\|}\, \exp\left\{\frac{-c^2}{\|\mathbf{v}\|^2}\right\}. \tag{7.62}$$

Without loss of generality, we will assume that $v_N = 1$.

$$\begin{aligned}\int \cdots \int_{\mathbb{R}^N} e^{-\mathbf{x}^T\mathbf{x}}\, \delta(\mathbf{v}^T\mathbf{x}-c)\, \mathrm{d}\mathbf{x} &= \int_{-\infty}^{\infty} \int \cdots \int_{\mathbb{R}^{N-1}} e^{-\mathbf{x}^T\mathbf{x}}\, \delta(\mathbf{v}^T\mathbf{x}-c)\, \mathrm{d}\mathbf{x}_l\, \mathrm{d}x_N \\ &= \int \cdots \int_{\mathbb{R}^{N-1}} e^{-(\mathbf{x}_l^T\mathbf{\Phi}\mathbf{x}_l+2\mathbf{x}_l^T\mathbf{w}+c^2)}\, \mathrm{d}\mathbf{x}_l,\end{aligned}$$

from Lemma 3.2.2. Note that relations (2.18) – (2.22) are valid. Since $\mathbf{\Phi}$ is positive definite (Lemma 3.2.2), we can write $\mathbf{\Phi} = \sqrt{\mathbf{\Phi}}^{\,T}\sqrt{\mathbf{\Phi}}$. Completing the square, and letting

$$\mathbf{y} = \sqrt{\mathbf{\Phi}}\, \mathbf{x}_l + \sqrt{\mathbf{\Phi}}^{\,-T}\mathbf{w},$$

we obtain

$$\int \cdots \int_{\mathbb{R}^{N-1}} e^{-(\mathbf{x}_l^T \mathbf{\Phi} \mathbf{x}_l + 2\mathbf{x}_l^T \mathbf{w} + c^2)} \, \mathrm{d}\mathbf{x}_l = \frac{1}{\sqrt{|\mathbf{\Phi}|}} e^{-c^2 + \mathbf{w}^T \mathbf{\Phi}^{-1} \mathbf{w}} \int \cdots \int_{\mathbb{R}^{N-1}} e^{-\mathbf{y}^T \mathbf{y}} \, \mathrm{d}\mathbf{y}$$
$$= \frac{\sqrt{\pi}^{N-1}}{\|\mathbf{v}\|} e^{-c^2 + \mathbf{w}^T \mathbf{\Phi}^{-1} \mathbf{w}}.$$

Equation (7.62) follows by noting that, by (2.18) – (2.22), we have

$$\begin{aligned} -c^2 + \mathbf{w}^T \mathbf{\Phi}^{-1} \mathbf{w} &= c^2(-1 + \mathbf{v}_l^T \mathbf{\Phi}^{-1} \mathbf{v}_l) \\ &= c^2(-1 + \mathbf{v}_l^T (\mathbf{I}_{N-1} - \frac{1}{\|\mathbf{v}\|^2} \mathbf{v}_l \mathbf{v}_l^T) \mathbf{v}_l) \\ &= c^2(-1 + \|\mathbf{v}_l\|^2 - \frac{\|\mathbf{v}_l\|^4}{\|\mathbf{v}\|^2}) \\ &= c^2(-1 + \frac{\|\mathbf{v}_l\|^2}{\|\mathbf{v}\|^2}) \\ &= \frac{-c^2}{\|\mathbf{v}\|^2}. \end{aligned}$$

We are now ready to prove the Lemma. Since $\mathbf{\Psi}$ is positive definite, write $\mathbf{\Psi} = \sqrt{\mathbf{\Psi}}^T \sqrt{\mathbf{\Psi}}$, and let $\mathbf{y} = \sqrt{\mathbf{\Psi}}\, \mathbf{x}$. Then we have

$$\int \cdots \int_{\mathbb{R}^N} e^{-\mathbf{x}^T \mathbf{\Psi} \mathbf{x}} \, \delta(\mathbf{v}^T \mathbf{x} - c) \, \mathrm{d}\mathbf{x} = \frac{1}{|\mathbf{\Psi}|^{1/2}} \int \cdots \int_{\mathbb{R}^N} e^{-\mathbf{y}^T \mathbf{y}} \, \delta\left((\sqrt{\mathbf{\Psi}}^{-T} \mathbf{v})^T \mathbf{y} - c \right) \, \mathrm{d}\mathbf{x}.$$

Noting that $\|(\sqrt{\mathbf{\Psi}}^{-T} \mathbf{v})\| = \sqrt{\mathbf{v}^T \mathbf{\Psi}^{-1} \mathbf{v}}$, the result follows from (7.62). □

Proof of Lemma 3.2.5

Proof. It is well known that

$$\int_0^\infty \frac{\sin tu}{u} \, \mathrm{d}u = \begin{cases} -\pi/2, & t < 0, \\ 0, & t = 0, \\ \pi/2, & t > 0. \end{cases}$$

Thus, we can define the unit step function $h(t)$ to be

$$\begin{aligned} h(t) &= \frac{1}{\pi} \int_0^\infty \frac{\sin tu}{u} \, \mathrm{d}u + \frac{1}{2} \\ &= \begin{cases} 0, & t < 0, \\ 1/2, & t = 0, \\ 1, & t > 0. \end{cases} \end{aligned}$$

We can then represent the delta function as the derivative of the unit step function, or

$$\begin{aligned}\delta(t) &= \lim_{\Delta\to 0} \frac{h(t+\frac{\Delta}{2}) - h(t-\frac{\Delta}{2})}{\Delta} \\ &= \left[\lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int_0^\infty \frac{\sin(t+\frac{\Delta}{2})u}{u} - \frac{\sin(t-\frac{\Delta}{2})u}{u}\ \mathrm{d}u\right].\end{aligned}$$

Proceeding formally, we derive

$$\begin{aligned}J &= \int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\mathbf{D}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \delta(\mathbf{x}^T\mathbf{x}-c)\ \mathrm{d}\mathbf{x} \\ &= \int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\mathbf{D}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \left[\lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int_0^\infty \frac{\sin(\mathbf{x}^T\mathbf{x}-c+\frac{\Delta}{2})u}{u} - \frac{\sin(\mathbf{x}^T\mathbf{x}-c-\frac{\Delta}{2})u}{u}\ \mathrm{d}u\right]\ \mathrm{d}\mathbf{x}.\end{aligned}$$

Interchanging the limit and the integral, we obtain

$$\begin{aligned}J &= \lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int\cdots\int_{\mathbb{R}^N} \int_0^\infty e^{-(\mathbf{x}^T\mathbf{D}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \left[\frac{\sin(\mathbf{x}^T\mathbf{x}-c+\frac{\Delta}{2})u}{u} - \frac{\sin(\mathbf{x}^T\mathbf{x}-c-\frac{\Delta}{2})u}{u}\right]\ \mathrm{d}u\ \mathrm{d}\mathbf{x} \\ &= \lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int_0^\infty \int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\mathbf{D}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \left[\frac{\sin(\mathbf{x}^T\mathbf{x}-c+\frac{\Delta}{2})u}{u} - \frac{\sin(\mathbf{x}^T\mathbf{x}-c-\frac{\Delta}{2})u}{u}\right]\ \mathrm{d}\mathbf{x}\ \mathrm{d}u \\ &= \lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int_0^\infty \frac{1}{u} \int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\mathbf{D}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \left[\sin(\mathbf{x}^T\mathbf{x}-c+\frac{\Delta}{2})u - \sin(\mathbf{x}^T\mathbf{x}-c-\frac{\Delta}{2})u\right]\ \mathrm{d}\mathbf{x}\ \mathrm{d}u \\ &= \lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int_0^\infty\ \Im\left\{\frac{1}{u} \int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\mathbf{D}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \left[e^{iu(\mathbf{x}^T\mathbf{x}-c+\frac{\Delta}{2})} - e^{iu(\mathbf{x}^T\mathbf{x}-c-\frac{\Delta}{2})}\right]\ \mathrm{d}\mathbf{x}\ \mathrm{d}u\right\},\end{aligned}$$

where $\Im\,(z)$ is the imaginary part of z. Combining the exponents in the integrand and interchanging the order of integration, we arrive at

$$\begin{aligned}J &= \lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int_0^\infty\ \Im\left\{\frac{1}{u} \int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T(\mathbf{D}-iu\mathbf{I})\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \left[e^{-iu(c-\frac{\Delta}{2})} - e^{-iu(c+\frac{\Delta}{2})}\right]\ \mathrm{d}\mathbf{x}\ \mathrm{d}u\right\} \\ &= \lim_{\Delta\to 0} \frac{1}{\pi\Delta} \int_0^\infty\ \Im\left\{\frac{1}{u}\left[e^{-iu(c-\frac{\Delta}{2})} - e^{-iu(c+\frac{\Delta}{2})}\right]\right\} \int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\tilde{\mathbf{D}}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \mathrm{d}\mathbf{x}\ \mathrm{d}u.\end{aligned}$$

Here $\tilde{\mathbf{D}} = \mathbf{D} - iu\mathbf{I}$, so that $\left|\tilde{\mathbf{D}}\right| = \prod_{j=1}^N (d_j - iu)$. Since $d_j > 0$, the inner integral becomes

$$\begin{aligned}\int\cdots\int_{\mathbb{R}^N} e^{-(\mathbf{x}^T\tilde{\mathbf{D}}\mathbf{x}+2\mathbf{v}^T\mathbf{x})}\ \mathrm{d}\mathbf{x} &= \frac{\sqrt{\pi}^{\,N}}{\left|\tilde{\mathbf{D}}\right|^{1/2}}\ e^{\mathbf{v}^T\tilde{\mathbf{D}}^{-1}\mathbf{v}} \\ &= \sqrt{\pi}^{\,N}\ \frac{e^{\sum_{j=1}^N \frac{v_j^2}{d_j-iu}}}{\prod_{j=1}^N \sqrt{d_j - iu}}.\end{aligned}$$

Plugging this into the expression for J, we obtain

$$
\begin{aligned}
I &= \pi^{\frac{N-2}{2}} \lim_{\Delta\to 0} \frac{1}{\Delta} \int_0^\infty \frac{1}{u}\, \Im\, \left[e^{-iu(c-\frac{\Delta}{2})} - e^{-iu(c+\frac{\Delta}{2})}\right] \frac{e^{\sum_{j=1}^N \frac{v_j^2}{d_j - iu}}}{\prod_{j=1}^N \sqrt{d_j - iu}}\, \mathrm{d}u \\
&= -\pi^{\frac{N-2}{2}} \frac{\mathrm{d}}{\mathrm{d}c} \int_0^\infty \frac{1}{u}\, \Im\, \left(e^{-iuc} \frac{e^{\sum_{j=1}^N \frac{v_j^2}{d_j - iu}}}{\prod_{j=1}^N \sqrt{d_j - iu}} \right) \mathrm{d}u \\
&= \pi^{\frac{N-2}{2}} \int_0^\infty \Re\, \left(e^{-iuc} \frac{e^{\sum_{j=1}^N \frac{v_j^2}{d_j - iu}}}{\prod_{j=1}^N \sqrt{d_j - iu}} \right) \mathrm{d}u \\
&= \pi^{\frac{N-2}{2}} \int_0^\infty \Re\, \left(\frac{e^{-iuc + \sum_{j=1}^N \frac{v_j^2(d_j+iu)}{d_j^2+u^2}}}{\prod_{j=1}^N \sqrt{d_j - iu}} \right) \mathrm{d}u,
\end{aligned}
$$

where $\Re(z)$ is the real part of z. Now represent $d_j + iu = r_j e^{i\theta_j}$, so that $r_j = \sqrt{d_j^2 + u^2}$ and $\theta_j = \tan^{-1} \frac{u}{d_j}$. Consequently, J may be rewritten in the following form:

$$
\begin{aligned}
J &= \pi^{\frac{N-2}{2}} \int_0^\infty \Re\, \frac{e^{-iuc + \sum_{j=1}^N \frac{v_j^2(d_j+iu)}{r_j^2}}}{\prod_{j=1}^N \sqrt{r_j}\; e^{\frac{-i\theta_j}{2}}}\, \mathrm{d}u \\
&= \pi^{\frac{N-2}{2}} \int_0^\infty \left(\prod_{j=1}^N \frac{1}{\sqrt{r_j}} \right) e^{\sum_{j=1}^N \frac{v_j^2 d_j}{r_j^2}} \cos\left\{ -cu + \sum_{j=1}^N \left[\frac{v_j^2 u}{r_j^2} + \frac{\theta_j}{2} \right] \right\} \mathrm{d}u \\
&= \pi^{\frac{N-2}{2}} \int_0^\infty \left(\prod_{j=1}^N (d_j^2 + u^2)^{-1/4} \right) e^{\sum_{j=1}^N \frac{v_j^2 d_j}{d_j^2+u^2}} \cos\left\{ -cu + \sum_{j=1}^N \left[\frac{v_j^2 u}{d_j^2 + u^2} + \frac{1}{2} \tan^{-1} \frac{u}{d_j} \right] \right\} \mathrm{d}u.
\end{aligned}
$$

The result follows by writing $\frac{v_j^2 d_j}{d_j^2+u^2} = \frac{v_j^2}{d_j} + \frac{v_j^2 u^2}{d_j(d_j^2+u^2)}$ in the exponent.

□

4

SIMULATIONS AND RESULTS

In order to assess the precision of the approach proposed above, we analyze the performance of the method using artificial and real data. In the artificial simulations, we test the method against both a normal and non-normal dataset. We then conduct a study with the Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset, which consists of high-dimensional radar images for a target/non-target classification.

For a given dataset, we compute the Maximum Entropy (3.32) and Hybrid (4.52) posteriors for each test vector. Since these decision rules are derived from a matrix-normal distribution, we will refer to them as the ME-MN (Maximum Entropy – Matrix Normal) and H-MN (Hybrid– Matrix Normal) methods, respectively. We examine the common measure of performance of classification algorithms, namely the percent of correct classifications, and we compare it with that of the linear SVM. As a decision rule for the approach considered in this paper, we assign the vector $\mathbf{z}$ to the class with the highest posterior probability. For the linear SVM we use a "one versus the rest" rule whenever $C > 2$. Note that since $n >> N$, the classes do not overlap and can be easily separated by a hyperplane.

4.1 Simulations with Normal Data

We first investigate how the proposed classification method works under ideal conditions, using data that matches the assumptions exactly, i.e. where the matrix $\mathbf{X}$ of training samples follows the matrix-variate normal distribution in (1.5) from Chapter 3. Using the definition of a matrix-variate normal random variable in terms of multivariate normal,

$$\mathbf{X} \sim N_{n,N}(\mathbf{M}\mathcal{E}^T, \mathbf{\Sigma} \otimes \mathbf{\Psi}) \iff \text{vec}(\mathbf{X}^T) \sim N_{nN}(\text{vec}(\mathcal{E}\mathbf{M}^T), \mathbf{\Sigma} \otimes \mathbf{\Psi}),$$

we generate the corresponding random vector and reshape it accordingly to generate the matrix of training samples. In choosing parameters $\mathbf{M}$, $\mathbf{\Sigma}$ and $\mathbf{\Psi}$ of (1.5), we note that the decision rules are determined only by vectors $\mathbf{m}_i,\ i = 1, \ldots, C$, matrix $\mathbf{\Psi}$ and the parameter σ^2. We choose the class means to be the canonical unit vectors in $\mathbb{R}^n$, i.e. $\mathbf{m}_i = \boldsymbol{\nu}_i$. Since $\mathbf{\Sigma}$ is not explicitly used in the decision rule, we choose $\mathbf{\Sigma}$ to be the identity matrix. In order to define the matrix $\mathbf{\Psi}$, recall that ideally it should have blocks of constants $\sigma_i^2,\ i = 1, \ldots, C$ along its diagonal. To construct such a matrix, we generate N vectors (whose elements are random uniformly distributed integers) $\mathbf{v}_{i,j},\ j = 1, \ldots, N_i,\ i = 1, \ldots, C$, and scale each vector so that $\|\mathbf{v}_{i,j}\|^2 = \sigma_i^2$. We then define $\sqrt{\mathbf{\Psi}} = [\mathbf{v}_{1,1} \cdots \mathbf{v}_{C,N_C}]$, so that $\mathbf{\Psi} = \sqrt{\mathbf{\Psi}}^{\,T}\sqrt{\mathbf{\Psi}}$ is a positive definite matrix with the desired diagonal. With these values for $\mathbf{M}, \mathbf{\Sigma}$, and $\mathbf{\Lambda}$, we generate the random matrix of training samples. Each test vector is generated from the $N(\mathbf{m}_i, \sigma_i^2\mathbf{\Sigma})$ distribution. Also, in all simulations, we use the uniform priors: $\pi_i = 1/C,\ i = 1, \ldots, C$, so that $N_1 = N_2 = \cdots N_C$.

In our simulation study we use $N_i = 5$ samples from each class available for training, $N_i^* = 100$ test vectors from each class to be classified, and the dimension of each vector is $n = 50$. In our first set of simulations, we consider the simple case where $C = 2$ and

$\sigma_1^2 = \sigma_2^2 = 0.2$. This value of σ_i^2 was chosen so that when the vectors are normalized and $\hat{\sigma}_i^2$ is estimated, it is close to the values for the MSTAR dataset. Figure 4.1 shows a comparison of the correct classification rates for ME-MN vs. SVM (top) and H-MN vs. SVM (bottom) for $M = 100$ simulation runs each.

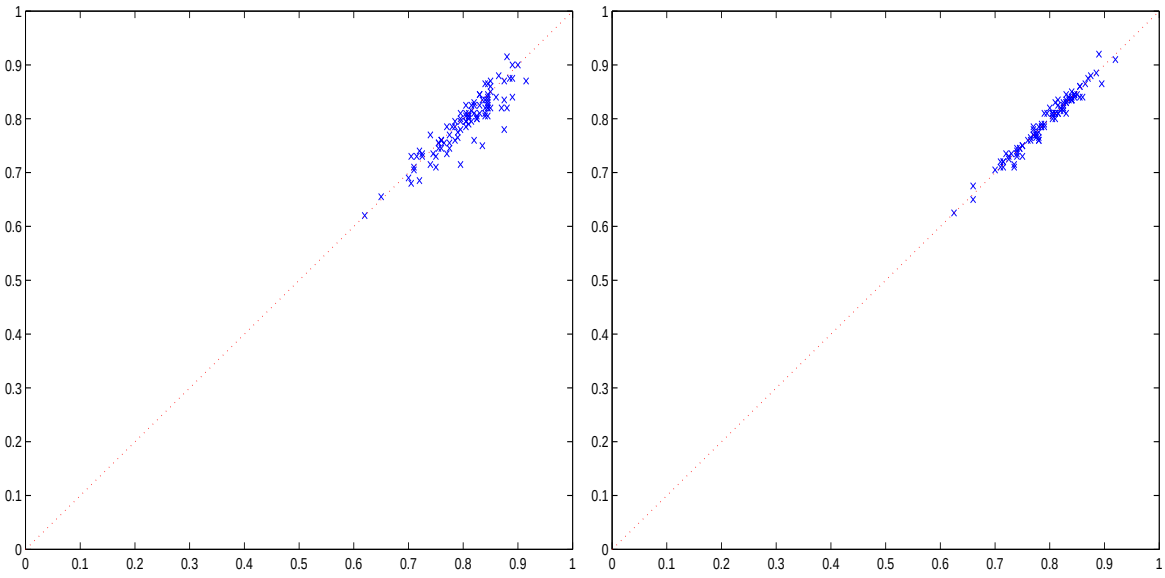

Figure 4.1: Percent Correct ME-MN vs. SVM (left) and H-MN vs. SVM (right) for Normal Data (100 iterations): $C = 2,\ n = 50,\ N_i = 5,\ N_i^* = 100,\ \sigma_1^2 = \sigma_2^2 = 0.2$

In Figure 4.1, each simulation run produces a point on the graph with coordinates $(x, y) \in (0, 1) \times (0, 1)$, where x is the percentage of vectors correctly classified using the new method, and y is the percentages correctly classified from the linear SVM. Hence, the point lies below the line $y = x$ if the new method is more precise than SVM and vice versa. Here we see that both methods yield classification rates that are very close to those of the linear SVM, although in this case the ME-MN exhibits a slight improvement over SVM. Furthermore, we notice that the H-MN method gives classification rates which are almost identical to SVM on *every* iteration.

Figure 4.2 shows the results for another two-class case, but now we choose $\sigma_1^2 = 0.2$ and

$\sigma_1^2 = 0.3$, and the remaining parameters are as in Figure 4.1. Like Figure 4.1, the top graph shows the classification rates of ME-MN vs. SVM, and the bottom H-MN vs. SVM. Once again, we see that both methods give classification rates which are very close to those of the linear SVM, with the H-MN method showing nearly identical rates for each iteration.

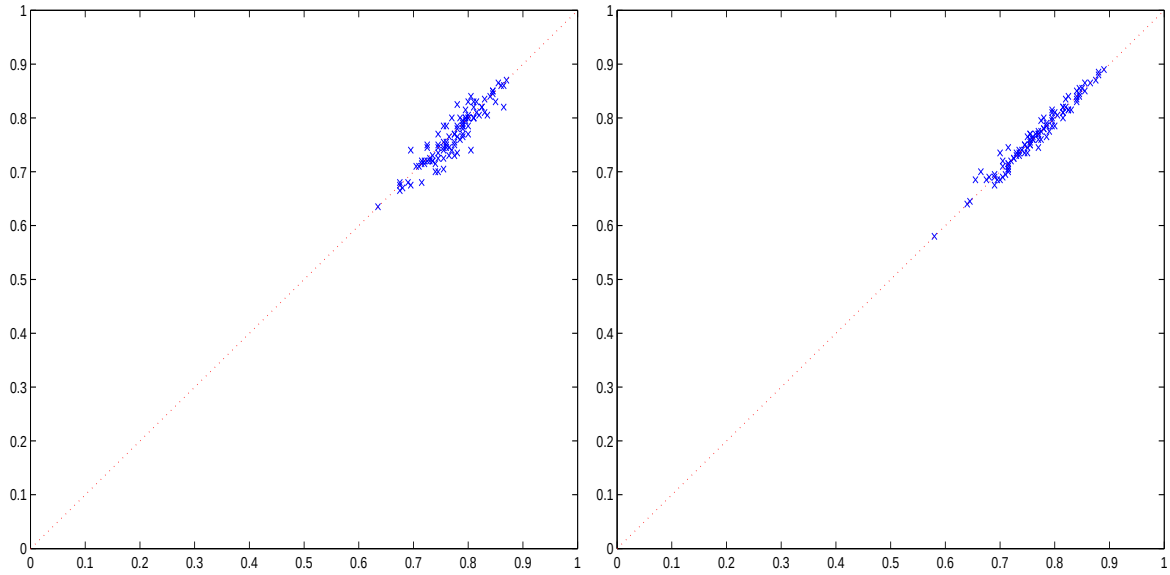

Figure 4.2: Percent Correct ME-MN vs. SVM (left) and H-MN vs. SVM (right) for Normal Data (100 iterations): $C = 2,\ n = 50,\ N_i = 5,\ N_i^* = 100,\ \sigma_1^2 = 0.2,\ \sigma_2^2 = 0.3$

Figure 4.3 shows the results for the case $C = 3$, $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.3$, and $\sigma_3^2 = 0.4$. The remaining parameters are the same as in Figures 4.1 and 4.2. Again we see classification rates for both methods that are very similar to SVM.

Note that Figures 4.1 – 4.3 do not show the actual correct classification percentages for each method. This is the purpose of Table 4.1, which gives a comparison of the average classification rates for the two methods vs. SVM for various parameters. Each row in the table corresponds to $M = 100$ runs, and the columns under "% Correct" contain the respective average correct classification rates over the M iterations.

One may also be interested in an accurate assessment of the posterior probabilities of

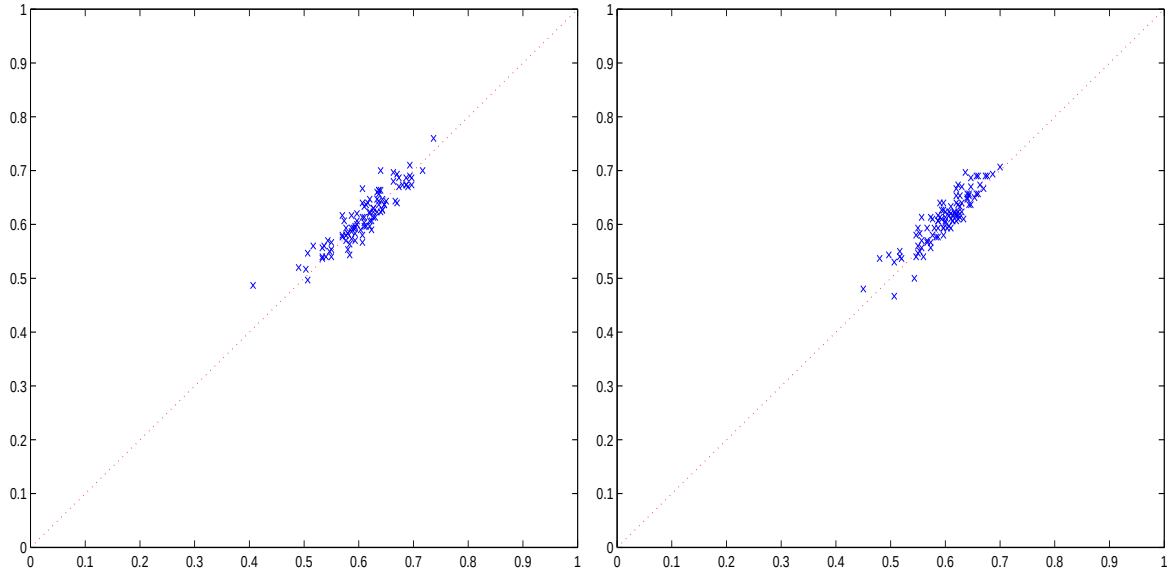

Figure 4.3: Percent Correct ME-MN vs. SVM (left) and H-MN vs. SVM (right) for Normal Data (100 iterations): $C = 3,\ n = 50,\ N_i = 5,\ N_i^* = 100,\ \sigma_1^2 = 0.2,\ \sigma_2^2 = 0.3,\ \sigma_3^2 = 0.4$

Table 4.1: Correct classification rates for Normal data averaged over M=100 iterations. Simulations are conducted with $N_i = 5$, $N_i^* = 100$ and $n = 50$.

parameters		% Correct		
C	σ_i^2	ME-MN	H-MN	SVM
2	(0.2, 0.2)	80.4	79.5	79.5
2	(0.2, 0.3)	77.3	76.6	76.7
3	(0.2, 0.3, 0.4)	61.0	60.0	61.1

each class which are provided by the ME-MN and H-MN methods. As a measure of their accuracy, we compare the computed posteriors with their empirical values. For this purpose, we discretized the interval $[0, 1]$ into bins of size Δ. In a single simulation run, we define the set $S_{i,k}$ to be the set of test vectors whose posteriors for class $\omega_i,\ i = 1, \ldots, C$, fall within the k-th interval, $[(k-1)\Delta, k\Delta],\ k = 1, ..., 1/\Delta$, and denote $\hat{S}_{i,k} = S_{i,k} \cap \omega_i$. Then, for the set $S_{i,k}$, the ratio $\#(\hat{S}_{i,k})/\#(S_{i,k})$ gives the percentage of them that are actually in class ω_i. Choosing $\Delta = 0.1$, we plot points with coordinates $x = (k - \frac{1}{2})\Delta$, $y = \text{ave}(\#(\hat{S}_{i,k})/\#(S_{i,k}))$, $k = 1, \ldots, 1/\Delta$, for each class $\omega_i,\ i = 1, \ldots, C$, averaged over M iterations. Figure 4.4 shows plots of this kind for the same set of simulations as in Figure 4.1, i.e. $C = 2,\ M = 100,\ n =$

50, $N_i = 5$, $N_i^* = 100$, $\sigma_1^2 = \sigma_2^2 = 0.2$. If the performance of the method were perfect and Δ were small enough, the graphs would coincide with the line $y = x$, indicating that the computed posteriors matched the true percentage exactly.

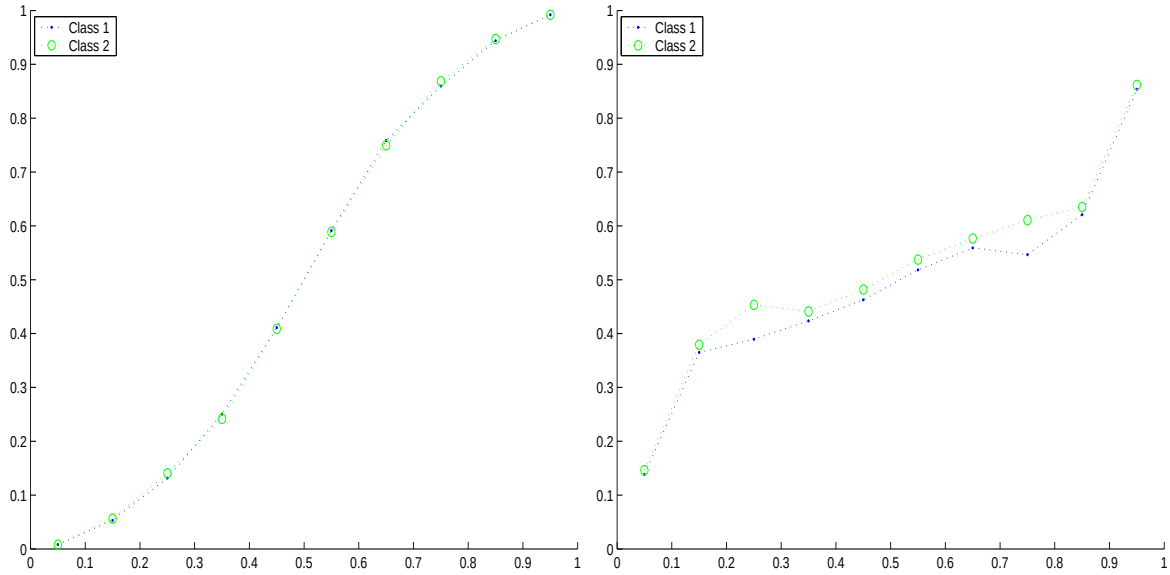

Figure 4.4: Posteriors vs. Empirical Values, ME-MN (left) and H-MN (right) for Normal Data (100 iterations): $C = 2$, $n = 50$, $N_i = 5$, $N_i^* = 100$, $\sigma_1^2 = \sigma_2^2 = 0.2$

In our example Δ is relatively large, however one can see that the graphs tend to follow the line $y = x$. Figure 4.5 demonstrates similar plots with the same parameter values as in Figure 4.2, i.e. $C = 2$, $M = 100$, $n = 50$, $N_i = 5$, $N_i^* = 100$, $\sigma_1^2 = 0.2$, $\sigma_2^2 = 0.3$.

In Figure 4.6, we show posteriors vs. their empirical values for $C = 3$, using the same parameters as in Figure 4.3, i.e. $M = 100$, $n = 50$, $N_i = 5$, $N_i^* = 100$, $\sigma_1^2 = 0.2, \sigma_2^2 = 0.3$, $\sigma_3^2 = 0.4$.

Note that when each class has a different value of σ_i^2, both methods tend to favor (assign more vectors to) the class with the smallest value of σ_i^2. Although it is not clear from the figures, the H-MN method tends to perform slightly better in the case of different σ_i^2, so for the MSTAR dataset we only study the H-MN method.

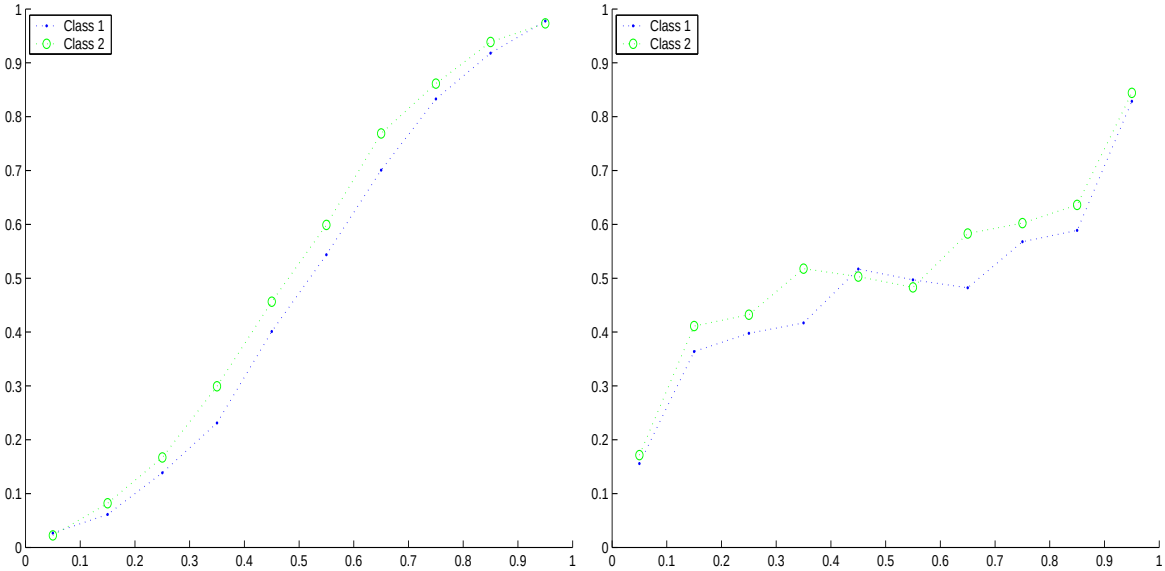

Figure 4.5: Posteriors vs. Empirical Values, ME-MN (left) and H-MN (right) for Normal Data (100 iterations): $C = 2,\ n = 50,\ N_i = 5,\ N_i^* = 100,\ \sigma_1^2 = 0.2,\ \sigma_2^2 = 0.3$

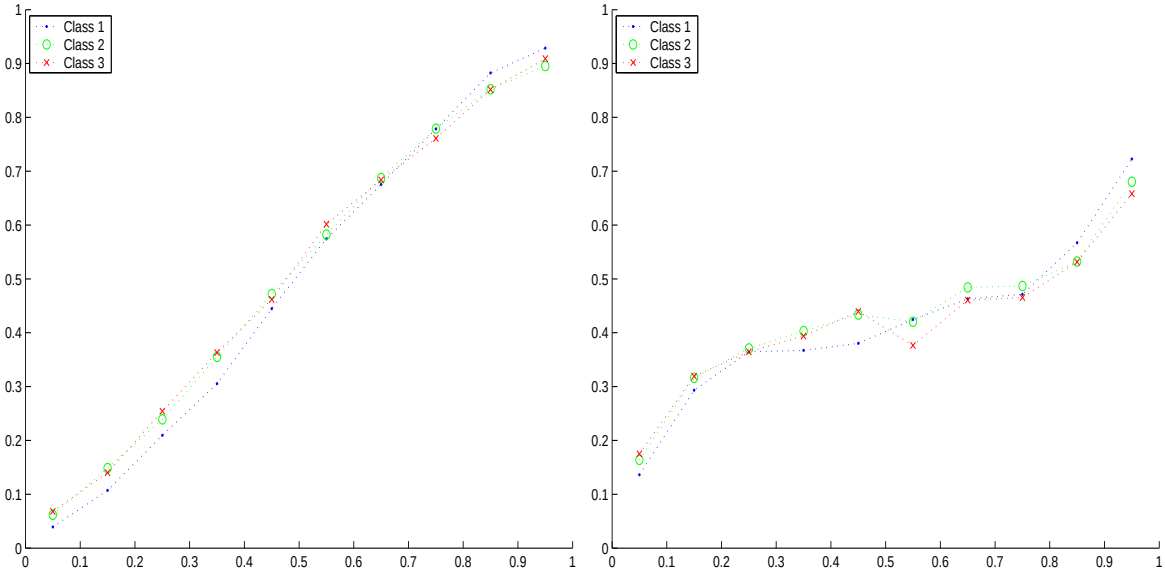

Figure 4.6: Posteriors vs. Empirical Values, ME-MN (left) and H-MN (right) for Normal Data (100 iterations): $C = 3,\ n = 50,\ N_i = 5,\ N_i^* = 100,\ \sigma_1^2 = 0.2,\ \sigma_2^2 = 0.3,\ \sigma_3^2 = 0.4$

4.2 Simulations with Non-Normal Data

The decision rules derived in this paper assume that the data is normally distributed. Since it is difficult to test the normality hypothesis in our setting (high-dimension, low sample

size), we instead test the ME-MN and ME-H methods against data which is not normally distributed. To do this, we use the Laplace (double exponential) distribution, which we denote by $L(\mu, \sigma^2)$. The pdf is given by

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2}\,\sigma} \exp\left\{ \frac{\sqrt{2}\,|x-\mu|}{\sigma} \right\}.$$

The vectors from class ω_i are generated so that if $t_{i,k}$ is the k^{th} element of a vector from class ω_i, then $t_{i,k} \sim L(\delta_{ik}, \sigma_i^2)$. Therefore, the class means are the canonical unit vectors, and there is independence between and within vectors. The goal in this study is to test not only deviations from normality, but also the affects of uncorrelated data. Table 4.2 shows the results of these simulations in a format identical to Table 4.1. The results again show similar classification rates as SVM, hence the methods do not appear to be sensitive to deviations from the assumptions.

Table 4.2: Correct classification rates for Laplacian data averaged over M=100 iterations. Simulations are conducted with $N_i = 5$, $N_i^* = 100$ and $n = 50$.

parameters		% Correct		
C	σ_i^2	ME-MN	H-MN	SVM
2	(0.2, 0.2)	80.0	80.4	79.8
2	(0.2, 0.3)	76.9	75.6	76.0
3	(0.2, 0.3, 0.4)	58.8	58.1	59.2

4.3 Application to Target Detection and Recognition

In real life situations one cannot expect that the assumptions of the model are satisfied exactly. For this reason, we study the performance of the proposed method using the public domain Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset, which is a collection of X-band synthetic aperture radar (SAR) images of 1 foot by 1 foot resolution. The dataset contains 40×40 pixel images of the T72 battle tank, the BMP armored personnel

carrier (APC), the BTR70 APC, and clutter. Images from the target class and clutter class are shown in Figure 4.7. For this dataset, we only show output for the ME-MN method, since it gave the most meaningful posterior probabilities. However, the H-MN method did provide classification rates nearly identical to SVM on every simulation run (like in Figures 4.1 and 4.2).

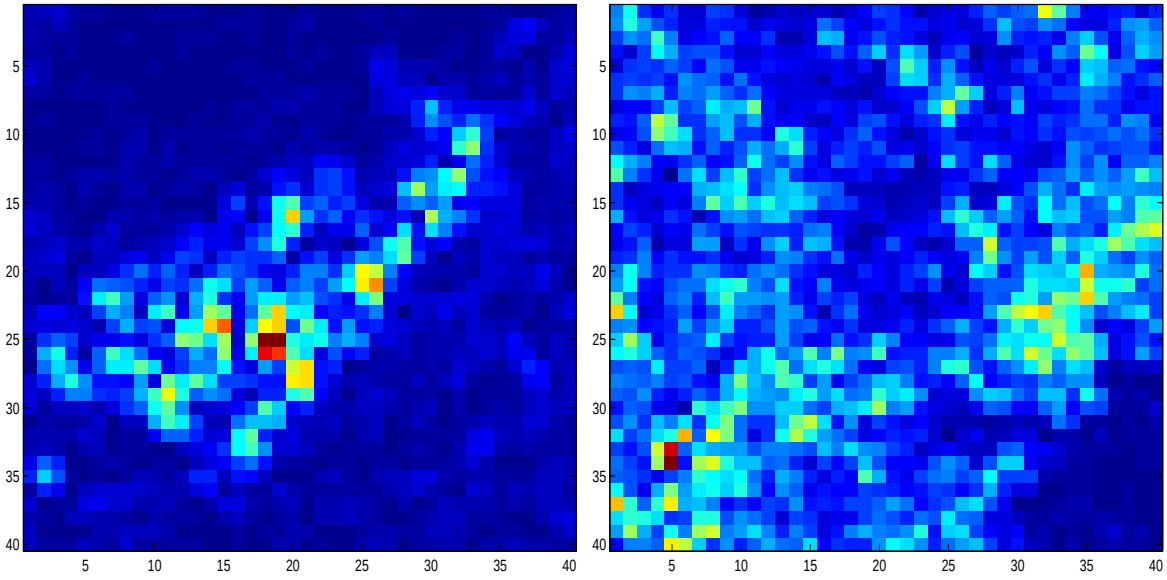

Figure 4.7: MSTAR Images of Target (left) and Clutter (right)

For the application of the ME-MN method, we cropped the images to 25×25 and scaled each vector to unit norm so that the estimate for $\mathbf{\Psi}$ has blocks of constants along the diagonal corresponding to σ_i^2. For a comparison to SVM, we perform 50 runs of simulations. In each of the runs, we pick up $N_i = 5,\ i = 1, 2$, images from each of the two classes as training samples, and another $N_i^* = 50,\ i = 1, 2$, vectors to be classified. In Figure 4.8, we represent the results of each simulation run as a point, similar to Figures 4.1 - 4.3. As one can see, the classification rates varied for both ME-MN and SVM, but on the average tended to be near the line $y = x$, confirming that the ME-MN method achieves classification precision similar to SVM – and in addition provides posterior probabilities for each class. A plot of

these posterior probabilities vs. their empirical values is given in Figure 4.9. In Figure 4.9, we see that the ME-MN method does provide meaningful posterior probabilities, while still achieving classification rates similar to the SVM.

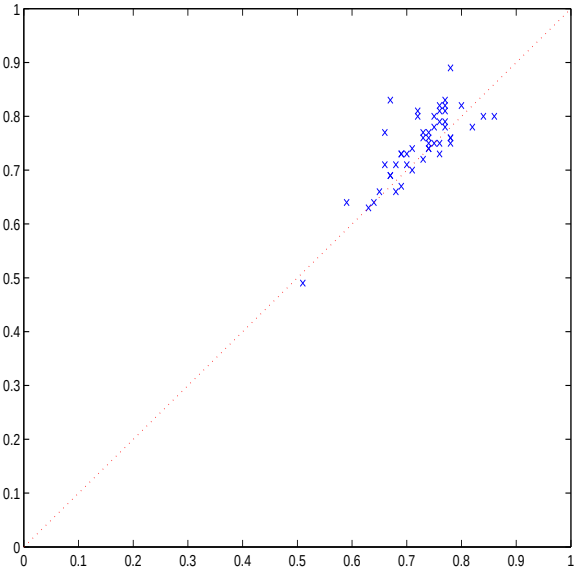

Figure 4.8: Percent Correct ME-MN vs. SVM for MSTAR Data (50 iterations): $C = 2,\ n = 625,\ N_i = 5,\ N_i^* = 50$

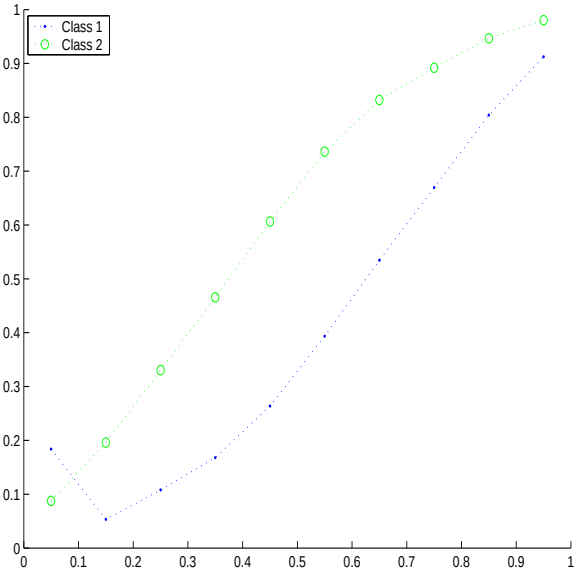

Figure 4.9: Posteriors vs. Empirical Values, ME-MN for MSTAR Data (50 iterations): $C = 2,\ n = 625,\ N_i = 5,\ N_i^* = 50$

4.4 Remarks

The method proposed in the paper is very efficient computationally and has very few limitations. It does not require much storage space apart from original matrix $\mathbf{X}$ and the new vector $\mathbf{z}$ to be classified, and all structures involved in the classification rule are of the small size $N \times N$. This distinguish our algorithm from classical decision rules which require the evaluation of the $n \times n$ matrix $\mathbf{\Sigma}$. In addition, both the Matrix Entropy and the Hybrid priors lead to a very efficient computation algorithm. After $\mathbf{X}^T\mathbf{X}$, $\mathbf{X}^T\mathbf{z}$ and $\mathbf{z}^T\mathbf{z}$ are evaluated, the computational complexity of the method is $O(N^2)$ where $N << n$, so it is more suitable than MCMC-based Bayesian classification [Mallick et al., 2005] in situations where the decision must be made in real time.

Our simulations show that our methods have misclassification rates very similar to SVM, but they also possesses convenient features of Bayesian approaches. In addition, we demonstrated that the posterior probabilities of the classes provided by our algorithm are reasonably close to their empirical values, and hence provide useful information in many real life situations.

5

CONCLUSIONS AND FUTURE WORK

In this dissertation, we introduced a new technique for the classification of high-dimensional vectors based on a small number of training samples. The method is based on the small vector $\mathbf{a}$, which can be interpreted as the coefficients of the projection of the observation $\mathbf{z}$ onto the space spanned by the training samples $L(\mathbf{X})$. The purpose of this method was to derive meaningful posterior probabilities while avoiding the "curse of dimensionality," which arises when trying to estimate the large covariance matrices in the decision theoretic approach.

This method employed the use of matrix-variate normal distributions, which have previously been used in classification only in the context of repeated measurements. Our assumption that the matrix of training samples follows a matrix-variate normal distribution can be viewed as an restriction to the method, since we did not perform any validation that the data was normally distributed. However, the normality assumption is difficult to test in this situation, since the majority of available tests are only suitable when n is small and $N \to \infty$. Even the normality tests which allow for $N < n$ [Liang et al., 2000, Tan et al., 2005] are not appropriate when $N << n$. Furthermore, they only test to reject normality, not confirm it. In this paper, we studied the effects of deviations from normality via simulations, and

the results confirmed that deviations from normality do not have a critical effect. Moreover, the normality assumption was only used to derive the conditions in Theorem 3.1.2, which established class-conditional relations on the coefficient vector $\mathbf{a}$.

Based on the conditions from Theorem 3.1.2, we chose the class-conditional distributions on the coefficient vector $\mathbf{a} \in \mathbb{R}^N$, where $N << n = \dim(\mathbf{z})$, thus avoiding the "curse of dimensionality." We proposed three different interpretations of the conditions – Delta, Maximum Entropy, and Hybrid – resulting in three different priors on $\mathbf{a}$, although the Delta prior proved to be difficult to implement.

The results of the numerical simulations were promising. For the simulated data, we observed that both the Maximum Entropy and Hybrid methods provided meaningful posterior probabilities, as evidenced by comparing them with their empirical counterparts. Furthermore, both methods provided classification rates very similar to the Linear Support Vector Machine. For the application to the target detection problem, we noticed that the Maximum Entropy method was closer to the true posterior probabilities, and the Hybrid method had a classification rate which matched the SVM rate almost exactly in every simulation run. Furthermore, we noticed that both methods tended to favor (assign more vectors to) the class with the smallest value of σ_i^2.

The work presented in this dissertation can be extended in several ways. First, classification based on the matrix $\hat{\mathbf{\Omega}} = \mathbf{X}^T\mathbf{X}$ can be extended to classification based on reproducing kernels $K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})\phi(\mathbf{y})$, for some function $\phi(\mathbf{x})$. Second, instead of arbitrary choices for the components d_1 and d_2 in matrix $\mathbf{\Lambda}$ (see equation (4.46) in Chapter 3), we can develop a meaningful algorithm which, for example, chooses d_1 and d_2 to maximize the marginal likelihood. We would then need to conduct simulations for this $\mathbf{\Lambda}$ and compare the results

with simulations when $\mathbf{\Lambda} = \sigma^2\mathbf{I}$. Finally, it would be a challenge to develop a fully Bayesian model of the SVM which is suitable when not necessarily all the training samples are involved in the classification process.

REFERENCES

J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Belmont, CA, 2001.

S. C. Choi. Classification of multiply observed data. *Biometrical Journal*, 14:8–11, 1972.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., New York, 2001.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 2000.

A. K. Gupta. On a classification rule for multiple measurements. *Computer and Mathematics with Applications*, 12A:301–308, 1986.

A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman and Hall CRC, Boca Raton, 2000.

F. S. Hillier and G. J. Lieberman. *Introduction to Operations Research*. McGraw-Hill Publishing, New York, 1986.

K. Jajuga, A. Sokolowski, and H. Bock, editors. *Classification, Clustering, and Data Analysis*. Springer, Berlin, 2001.

J. P. Keener. *Principles of Applied Mathematics*. Perseus Books, Cambridge, MA, 2000.

W. J. Krzanowski and F. H. Marriott. *Multivariate Analysis: Part 2, Classification, Covariance Structures & Repeated Measurements*. Edward Arnold, London, 1995.

Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines: Theory and applications to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

J. Liang, R. Li, H. Fang, and K. Fang. Testing multinormality based on low-dimensional projection. *Journal of Statistical Planning and Inference*, 86:129–141, 2000.

Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6:259–275, 2002.

B. K. Mallick, D. Ghosh, and M. Ghosh. Bayesian classification of tumors using gene expression data. *Journal of the Royal Statistical Society*, 2005. to appear.

A. M. Mathai. *Jacobians of Matrix Transformations and Functions of Matrix Argument*. World Scientific, Singapore, 1997.

K. S. Miller. *Multidimensional Gaussian Distributions*. John Wiley & Sons, Inc., New York, 1964.

B. Muise. *Quadratic Filters for Automatic Pattern Recognition.* PhD thesis, University of Central Florida, Orlando, FL, 2003.

J. M. Ortega. *Matrix Theory.* Plenum Press, New York, 1987.

M. Tan, H. Fang, G. Tian, and G. Wei. Testing multivariate normality in incomplete data of small sample size. *Journal of Multivariate Analysis*, 93:164179, 2005.

A. B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32(1):135–166, 2004.

V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, Inc., New York, 1998.

G. Wahba. Soft and hard classification by reproducing kernel hilbert space methods. *Proccedings of the National Academy of Sciences*, 99:14524–16530, 2002.

A. R. Webb. *Statistical Pattern Recognition.* John Wiley & Sons, Inc., West Sussex, England, 2002.

R. Yang and J. O. Berger. Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, 22(3):1195–1211, 1994.