

**Analysis and Design of Linear Classifiers for  
High-Dimensional, Small Sample Size Data Using  
Asymptotic Random Matrix Theory**

Dissertation by

Lama B. Niyazi

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy


King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

©November, 2023

Lama B. Niyazi

All rights reserved

 <https://orcid.org/0000-0003-0390-8332>

**ABSTRACT**

Analysis and Design of Linear Classifiers for High-Dimensional,  
Small Sample Size Data Using Asymptotic Random Matrix Theory

Lama B. Niyazi

November, 2023

Due to a variety of potential barriers to sample acquisition, many of the datasets encountered in important classification applications, ranging from tumor identification to facial recognition, are characterized by small samples of high-dimensional data. In such situations, linear classifiers are popular as they have less risk of overfitting while being faster and more interpretable than non-linear classifiers. They are also easier to understand and implement for the inexperienced practitioner.

In this dissertation, several gaps in the literature regarding the analysis and design of linear classifiers for high-dimensional data are addressed using tools from the field of asymptotic Random Matrix Theory (RMT) which facilitate the derivation of limits of relevant quantities or distributions, such as the probability of misclassification of a particular classifier or the asymptotic distribution of its discriminant, in the RMT regime where both the sample size and dimensionality of the data grow together. The resulting insights extracted from these limits allow for a deeper understanding of the classifier's behavior as well as lay out the groundwork from which to propose modifications to the classifier in order to improve its performance. Asymptotic RMT is also used in this dissertation to derive estimators of quantities of interest which are consistent in the RMT regime. Besides this, the estimators facilitate the tuning of classifier hyperparameters without resort to empirical methods such as cross-validation which can be very computationally-taxing when high-dimensional data is involved.

This work begins with an asymptotic study of the discriminant-averaging and vote-averaging Randomly-Projected Linear Discriminant Analysis (RP-LDA) ensemble classifiers – two high-dimensional variants of the classical Linear Discriminant Analysis (LDA) classifier based on random projections. The asymptotically optimal ensemble based on randomly-projected LDA discriminants for Gaussian data is found to be a form of discriminant-averaging and it is shown that selecting projections for inclusion in the ensemble based on some metric of expected performance offers no performance advantage. Furthermore, a closer look at the infinite ensemble version of the discriminant-averaging RP-LDA ensemble classifier, where the Marzetta estimator of the precision matrix arises in the discriminant, reveals that the Marzetta estimator behaves as an inversion of a linear regularization of the sample covariance matrix. This has the implication that the discriminant-averaging RP-LDA ensemble classifier asymptotically behaves as a special case of Regularized Linear Discriminant Analysis (R-LDA) with coarser parameter tuning since its regularization parameter varies with the integer projection dimension. From there, the class of rotationally-invariant estimators–to which the Marzetta estimator belongs–is studied in the context of classification. A modified LDA classifier based on a rotationally-invariant estimator of the sample covariance matrix having non-linear shrinkage which minimizes the probability of misclassification is proposed. Finally, a technique for tuning the weight vector of a generic binary linear classifier is developed. This technique is shown to not only yield performance gains for LDA in a small sample scenario, but to also compensate for the performance loss due to non-optimal native hyperparameters of classifiers such as the Support Vector Machine (SVM) with linear kernel, which would otherwise be computationally costly to tune optimally. The dissertation is concluded with an application of the weight vector tuning technique to the transfer learning of a deep neural network, showcasing the ubiquity of linear classifiers and the potential for widespread applicability of the proposed technique.

اللّهُمَّ مِنْكَ وَإِلَيْكَ

## ACKNOWLEDGEMENTS

This dissertation is a culmination of the efforts of many people over many years to whom I wish to express my gratitude. Firstly, I want to thank my advisors, Prof. Mohamed-Slim Alouini and Prof. Tareq Al-Naffouri, for supporting me, advising me, and providing me with many opportunities to learn and develop throughout my time at KAUST. I am also grateful to Prof. Hayssam Dahrouj who was my professor while I was an undergraduate, as well as my mentor throughout my graduate studies. He has always supported me and advocated for me, which I deeply appreciate. To my other mentor, Dr. Abla Kammoun, I am especially thankful for teaching me everything I know about random matrix theory. I am grateful for her patience and kindness over the years and greatly admire her dedication. I also want to thank my committee members, Prof. Basem Shihada and Prof. Yehia Massoud, and the external examiners, Prof. Babak Hassibi and Prof. K.V.S. Hari, for their valuable time and feedback.

I am thankful for all of my friends, especially Sarah Toonsi, Wafa Hedhly, Sondos Shanti, and Sara Helal, who kept me company over the years. I am greatly indebted to my parents who sacrificed a lot for me to be able to get to this point. I am especially grateful to my father who has always encouraged me to pursue learning for its own sake. I cannot express my appreciation for my parents and my siblings in words.

Finally, I want to acknowledge KAUST for providing a wonderful environment where I got the opportunity to learn from many passionate people and work on world-class research. Of course none of this would have been possible without the help and blessings of Allah Almighty.

## TABLE OF CONTENTS

<b>Abstract</b>	<b>2</b>
<b>List of Abbreviations</b>	<b>10</b>
<b>Notation</b>	<b>12</b>
<b>List of Figures</b>	<b>13</b>
<b>List of Tables</b>	<b>16</b>
<b>Examination Committee Page</b>	<b>17</b>
<b>1 Introduction</b>	<b>18</b>
1.1 Motivation . . . . .	18
1.1.1 Why high-dimensional, small sample size data? . . . . .	18
1.1.2 Why linear classifiers? . . . . .	20
1.1.3 Why asymptotic RMT? . . . . .	21
1.2 RMT . . . . .	22
1.3 Overview of work and contributions . . . . .	26
<b>2 Technical Background</b>	<b>29</b>
2.1 Linear classifiers . . . . .	29
2.2 Classification setting . . . . .	29
2.3 LDA . . . . .	30
2.3.1 The Bayes classifier . . . . .	31
2.3.2 The LDA classifier in practice . . . . .	32
2.3.3 Additional optimality properties . . . . .	32
2.4 RMT concepts . . . . .	34
2.4.1 Basic definitions . . . . .	34
2.4.2 Frequently-used lemmas and results . . . . .	42
<b>3 RP-LDA Ensembles</b>	<b>44</b>
3.1 Background . . . . .	45
3.1.1 Random projections . . . . .	45
3.1.2 RP-LDA ensemble classifiers . . . . .	47
3.2 Contributions . . . . .	48

3.3	Classifier definitions . . . . .	50
3.3.1	The single RP-LDA classifier . . . . .	51
3.3.2	RP-LDA ensemble classifiers . . . . .	52
3.4	Asymptotic insights . . . . .	55
3.4.1	Convergence of discriminant statistics and asymptotic distributions . . . . .	56
3.4.2	Asymptotically optimal ensemble of RP-LDA discriminants	61
3.5	Turning theory into practice . . . . .	72
3.5.1	G-estimators . . . . .	73
3.5.2	Tuning the discriminant-averaging RP-LDA ensemble parameters . . . . .	79
3.6	The discriminant-averaging RP-LDA infinite ensemble as a special case of R-LDA . . . . .	85
<b>4</b>	<b>General Shrinkage for LDA Classification</b>	<b>88</b>
4.1	Background . . . . .	88
4.2	Contributions . . . . .	91
4.3	Rotationally-invariant estimators in the context of LDA classification	93
4.3.1	The rotationally-invariant LDA Rule . . . . .	93
4.3.2	Bayes shrinkage . . . . .	94
4.4	Main results . . . . .	95
4.4.1	Proposed form of shrinkage . . . . .	95
4.4.2	Assumptions and main results . . . . .	97
4.5	Simulations . . . . .	99
<b>5</b>	<b>Weight-Vector Tuning of Linear Classifiers</b>	<b>106</b>
5.1	Background . . . . .	107
5.2	Contributions . . . . .	108
5.3	Weight vector tuning procedure . . . . .	110
5.3.1	Known class means . . . . .	111
5.3.2	Unknown class means . . . . .	117
5.4	Asymptotic analysis and tuning of the parameterized LDA classifier	126
5.4.1	Asymptotic analysis . . . . .	127
5.4.2	Tuning the parameterized LDA classifier . . . . .	133
5.5	Transfer learning application . . . . .	135
5.5.1	Simulation setup . . . . .	136
5.5.2	Results . . . . .	138
<b>6</b>	<b>Concluding Remarks</b>	<b>142</b>

<b>References</b>	<b>143</b>
<b>Appendices</b>	<b>150</b>
<b>A Proofs for Chapter 3</b>	<b>151</b>
<b>B Proofs for Chapter 4</b>	<b>185</b>
<b>C Proofs for Chapter 5</b>	<b>207</b>



## LIST OF ABBREVIATIONS

a.s.	almost surely
AUC	Area Under the Curve
CDF	Cumulative Distribution Function
CV	Cross-Validation
DCT	Discrete Cosine Transform
DE	Deterministic Equivalent
ESD	Empirical Spectral Distribution
FNR	False Negative Rate
FPR	False Positive Rate
FTIR	Fourier Transform Infrared Spectroscopy
i.i.d.	independent and identically distributed
LDA	Linear Discriminant Analysis
LSD	Limit Spectral Distribution
MAP	Maximum A Posteriori
MMSE	Minimum Mean Square Error
MSE	Mean Square Error
NPV	Negative Predictive Value
PCA	Principal Components Analysis
PDF	Probability Density Function
PMF	Probability Mass Function
PPV	Positive Predictive Value
PRIAL	Percentage Relative Improvement in Average Loss
PRIE	Percentage Relative Improvement in Error
QIS	Quadratic Inverse Shrinkage
R-LDA	Regularized Linear Discriminant Analysis
RMT	Random Matrix Theory
ROC	Receiver Operating Characteristic
RP-LDA	Randomly-Projected Linear Discriminant Analysis
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TNR	True Negative Rate

TPR	True Positive Rate		
WS	Weighted Shrinkage		
WS-LDA	Weighted-Shrinkage	Linear	Discriminant
	Analysis		

## NOTATION

$(x)^+$	The maximum between $x \in \mathbb{R}$ and 0
$Q(\cdot)$	Standard Gaussian complementary Cumulative Distribution Function (CDF)
$X \sim Y$	$X$ is distributed as $Y$
$\Phi(\cdot)$	Standard Gaussian CDF
$\delta(x)$	The dirac delta located at zero
$\mathbb{E}[X]$	Expectation of $X$
$\mathbb{1}_{\mathcal{A}}(x)$	Indicator function of the set $\mathcal{A}$ which is equal to 1 when $x \in \mathcal{A}$ and to 0 otherwise
$\mathcal{CB}(n, p, \rho)$	Correlated binomial distribution with $n$ trials each having a probability of success of $p$ and correlated to each other by correlation coefficient $\rho$
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean $\mu$ and variance $\sigma^2$
$\mathbf{0}_p$	The all-zeros $p \times 1$ vector
$\mathbf{1}_p$	The all-ones $p \times 1$ vector
$\mathbf{A}_p$	$p \times p$ square matrix
$\mathbf{I}_p$	The $p \times p$ identity matrix
$\xrightarrow{\text{a.s.}}$	Convergence almost-surely
$\xrightarrow{\text{d}}$	Convergence in distribution
$\xrightarrow{\text{p}}$	Convergence in probability
$a \asymp b$	$a - b \xrightarrow{\text{a.s.}} 0$

## LIST OF FIGURES

2.1	Comparison of the histogram of a Wigner matrix ( $p = 1500$ ) with the semi-circle law . . . . .	36
2.2	Comparison of the eigenvalue plot of a non-Hermitian matrix ( $p = 1000$ ) with the full-circle law . . . . .	37
2.3	Comparison of the histogram of a Wishart matrix ( $p = 1600, n = 4000$ ) with the Marchenko-Pastur law . . . . .	38
3.1	Class-conditional asymptotic distributions of the discriminant-averaging ensemble $M = 10$ . . . . .	59
3.2	Class-conditional asymptotic distributions of the discriminant-averaging ensemble $M = 1, M = 10$ , and $M = \infty$ . . . . .	59
3.3	ROCs of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on Gaussian mixture model data. . . . .	66
3.4	Testing error of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on Gaussian mixture model data. . . . .	67
3.5	ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘gastro_WL’ dataset. . . . .	71
3.6	ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘gastro_NB’ dataset. . . . .	72
3.7	ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘prostate’ dataset. . . . .	73
3.8	Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘colon’ dataset. . .	74
3.9	Iterated 10-fold CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘leukemia_big’ dataset.	75

3.10	Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘leukemia_small’ dataset.	76
3.11	Iterated 10-fold nested CV estimate of the error rate of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘prostate’ dataset.	77
3.12	Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘prostate_full’ dataset.	78
3.13	Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘phoneme_aa_ao’ dataset.	79
3.14	Ratio of infinite to finite discriminant-averaging RP-LDA ensemble classifier error on Gaussian mixture model data. . . . .	81
4.1	Plot of the PRIE against varying data dimension, $p$ , where $p$ and $n$ grow together at a fixed ratio of $1/3$ . . . . .	103
4.2	Plot of the PRIE against varying concentration ratio, $p/n$ . . . . .	103
4.3	Plot of the PRIE against varying condition number, $\theta$ . . . . .	104
4.4	Plot of the PRIE against varying class $\mathcal{C}_0$ prior, $\pi_0$ . . . . .	105
5.1	Plot of the expected testing error of the modified discriminant against $\alpha$ for a randomly generated weight vector $\mathbf{w}$ . . . . .	117
5.2	Plots of expected testing error averaged over 100 training sets for data generated from classes with a common $\Sigma$ . Here, $p = 400$ and $n = 450$ . . . . .	121
5.3	Plots of expected testing error averaged over 100 training sets for data generated from classes with a common $\Sigma$ . Here, $p = 10$ and $n = 500$ . . . . .	121
5.4	Plots of expected testing error averaged over 100 training sets for data generated from classes with a common $\Sigma$ . Here, $p = 300$ and $n = 100$ . . . . .	122
5.5	Plots of expected testing error averaged over 100 training sets for data generated from classes with distinct $\Sigma_0$ and $\Sigma_1$ . Here, $p = 400$ and $n = 450$ . . . . .	123
5.6	Plots of expected testing error averaged over 100 training sets for data generated from classes with distinct $\Sigma_0$ and $\Sigma_1$ . Here, $p = 300$ and $n = 100$ . . . . .	123

5.7	Plot of expected testing error of $\alpha$ -SVM with penalty set to 1 averaged over 100 training sets for data generated from classes with distinct $\Sigma_0$ and $\Sigma_1$ . Here, $p = 400$ and $n = 450$ . . . . .	124
5.8	Plots of testing error on USPS digit pairs of $\alpha$ -SVM with penalty set non-optimally . . . . .	126
5.9	Plots of testing error estimates of classifying USPS digit pairs for LDA, the nearest centroid, and $\alpha$ -LDA as well as the G-estimator $\hat{\epsilon}$ of the $\alpha$ -LDA expected testing error. . . . .	133
5.10	Plots of testing error estimates of classifying phonemes ‘aa’ and ‘ao’ for LDA, the nearest centroid, and $\alpha$ -LDA as well as the G-estimator $\hat{\epsilon}$ of the $\alpha$ -LDA expected testing error. . . . .	135
5.11	Plot of testing error rate of the weight vector tuned GoogLeNet which was transfer learned on ‘daisy’ and ‘sunflower’ images. . . . .	139
5.12	Plot of testing error rate of the weight vector tuned GoogLeNet which was transfer learned on ‘rose’ and ‘tulip’ images. . . . .	139
5.13	Plot of testing error rate of the weight vector tuned GoogLeNet which was transfer learned on ‘pizza’ and ‘hamburger’ images. . . . .	139
5.14	Plot of testing error rate of the weight vector tuned ResNet-50 which was transfer learned on ‘daisy’ and ‘sunflower’ images. . . . .	140
5.15	Plot of testing error rate of the weight vector tuned ResNet-50 which was transfer learned on ‘rose’ and ‘tulip’ images. . . . .	140
5.16	Plot of testing error rate of the weight vector tuned ResNet-50 which was transfer learned on ‘pizza’ and ‘hamburger’ images. . . . .	140

## LIST OF TABLES

1.1	Table of benchmark gene microarray datasets, number of their features, sample size, and classes. . . . .	19
1.2	Table of benchmark face recognition datasets, number of their features (expressed as a resolution), and their sample size. . . . .	19
3.1	Datasets and their properties . . . . .	68
3.2	AUCs corresponding to the ROCs of the discriminant-averaging, discriminant-averaging-with-selection, vote-averaging, and vote-averaging-with-selection ensembles applied to real data. . . . .	70
3.3	Table of average testing errors and parameter settings of the infinite discriminant-averaging ensemble classifier, where $d$ is tuned based on cross-validation, and the finite discriminant-averaging ensemble classifier where $M$ and $d$ are tuned by Algorithm 1 based on cross-validation, G-estimators, and the heuristic on Gaussian mixture model data. Here $\psi = 0.95$ . For comparison, the Bayes error is 0.0401. . . . .	83
3.4	Table of average testing errors and parameter settings of the infinite discriminant-averaging ensemble classifier, where $d$ is tuned based on cross-validation, and the finite discriminant-averaging ensemble classifier where $M$ and $d$ are tuned by Algorithm 1 based on cross-validation, G-estimators, and the heuristic on the ‘phoneme_aa_ao’ dataset. Here $\psi = 0.99$ . . . . .	84

## EXAMINATION COMMITTEE PAGE

The dissertation of Lama B. Niyazi is approved by the examination committee.

Committee Chairperson: Mohamed-Slim Alouini

Committee Co-Chair: Tareq Y. Al-Naffouri

Committee Members: Abla Kammoun, Hayssam Dahrouj, Basem Shihada, Yehia Massoud, Babak Hassibi, and K.V.S. Hari

# Chapter 1

## Introduction

### 1.1 Motivation

This dissertation is concerned with the analysis and design of linear methods for the classification of high-dimensional, small sample size data using tools from asymptotic RMT. In order to motivate this study, this section answers three primary questions:

1. Why high-dimensional, small sample size data?
2. Why linear classifiers?
3. Why asymptotic RMT?

#### 1.1.1 Why high-dimensional, small sample size data?

Table 1.1 lists various benchmark datasets obtained from gene microarray experiments along with their dimensionality and sample size. Among other tasks, these datasets are commonly used for identification of tumor samples as being malignant or benign. Comparing the ‘Features’ column to the ‘Samples’ column for each dataset it is clear that the data dimensionality far exceeds the sample size, with the number of features typically being on the order of thousands while the sample size is on the order of hundreds.

Table 1.2 lists several benchmark face recognition datasets along with their image resolutions and the number of images per subject (where an individual subject represents a class). Considering just the entry corresponding to the ORL Database of Faces in the first row, which is grayscale, there are  $92 \times 112 = 10304$

Dataset	Features	Samples	Classes	Reference
Colon	2000	57	‘Tumor’ (37) or ‘Normal’ (20)	(Alon et al., 1999)
Ovarian	15154	253	‘Cancer’ (162) or ‘Non-cancer’ (91)	(Petricoin et al., 2002)
Prostate	12600	102	‘Tumor’ (52) or ‘Normal’ (50)	(Singh et al., 2002)
Leukemia	7129	72	‘ALL’ (47) or ‘AML’ (25)	(Golub et al., 1999)
ALL	12625	79	‘NEG’ (42) or ‘BCR/ABL’ (37)	(Chiaretti et al., 2004)

Table 1.1: Table of benchmark gene microarray datasets, number of their features, sample size, and classes.

Dataset	Features	Samples	Reference
ORL	$92 \times 112 \times 1$	10 images $\times$ 40 subjects	(Samaria and Harter, 1994)
YALE	$320 \times 243 \times 1$	11 images $\times$ 15 subjects	(Belhumeur et al., 1997)
AR	$576 \times 768 \times 3$	26 images $\times$ 126 subjects	(Martinez and Benavente, 1998)
UMIST	$220 \times 220 \times 1$	19-36 images $\times$ 20 subjects	(Graham and Allinson, 1998)
JAFFE	$256 \times 256 \times 1$	7 images $\times$ 10 subjects	(Lyons et al., 1998)

Table 1.2: Table of benchmark face recognition datasets, number of their features (expressed as a resolution), and their sample size.

features per image with a total of 400 samples of just 10 samples per class. For color images, the features are tripled. Again, the number of features in this data greatly outnumbers its sample size.

The above showcases just two of the domains where such data is encountered. With advances in technology, high throughput data has become common in many fields. Data acquisition, however, may be restricted due to costs, difficulty of obtaining patient consent, and rarity of the desired class or phenotype, naturally leading to the occurrence of high-dimensional, small sample size data. For example, technologies such as mass spectrometry, Fourier Transform Infrared Spectroscopy (FTIR), chromatography, and nuclear magnetic resonance, can result in massive spectral data, yet many chemometric datasets are lacking in sample size. A study of FTIRs for the classification of ionic liquids as exhibiting high or low antibacterial activity consisted of only 36 samples of 1676 features each (Mehmood and Iqbal, 2021). Likewise, an FTIR study of coffee for species classification consisted of only 56 samples of 286 features each (Kanwal et al., 2021). Small sample size and high-dimensionality is also common in portfolio optimization data where there are many assets which must be optimized based

on short-term returns data (Liao, 2019), and as shown in Tables 1.1 and 1.2 it is also the norm in gene microarray and face recognition datasets. Thus the answer to the question of why we consider such data is that it is everywhere and that standard classification methods tend to fail in dealing with it as is explained in the next section.

### 1.1.2 Why linear classifiers?

Classical machine learning methods such as the k-nearest neighbors, SVMs, linear regression, Markowitz portfolio selection (Ledoit and Wolf, 2022a), and discriminant analysis methods are designed with low-dimensional data in mind, in the sense that they rely on low-dimensional notions of distance and/or statistical results which hold for large samples of low-dimensional data. As a rule-of-thumb, the sample size must be at least 100 times the dimensionality in order for such methods to work (Liao, 2019). However, as described in the previous section, in many modern applications, it is common to encounter very high-dimensional data where the data dimensionality may be significantly larger than the number of samples of the data.

Linear classifiers offer some hope to escape the effects of the curse of dimensionality when dealing with high-dimensional, small sample data. They are less likely to overfit (Allison et al., 2006), faster to train and test, and easier to understand and implement than more complex methods. The literature attests to the advantage and suitability of linear classifiers over non-linear classifiers for high-dimensional, small sample size data as many of the classifiers designed for such data are linear. References (Tong et al., 2020), (Sifaou et al., 2020), (Huang et al., 2010), (Li et al., 2022), (Durrant and Kabán, 2015), (Cannings and Samworth, 2017), (Lu and Qiao, 2018), (Witten and Tibshirani, 2011), and (Chatterjee et al., 2023) all propose linear classifiers for general high-dimensional, small sample size data. References (Golub et al., 1999), (Tibshirani et al., 2002), (Xu et al., 2009), (Huynh et al., 2018), and (Mehmood et al., 2022) propose lin-

ear classifiers specifically for gene microarray data, while (Zou and Hastie, 2005) and (Huynh et al., 2019) propose feature selection methods combined with linear classifiers for gene microarray data. Hua et al. (2009) use linear SVM and LDA as base classifiers with which to test the performance of a number of feature selection methods for gene microarray data. The Fisherface technique is a popular method widely used in face recognition where the data is preprocessed by Principal Components Analysis (PCA) before applying the linear LDA for classification (Sharma and Paliwal, 2015). Also related to face recognition, references (Nakouri, 2021), (Thomaz et al., 2006), (Song et al., 2004), (Song et al., 2007), and (Murtaza et al., 2014) develop linear classifiers to address the high-dimensionality and small sample size of such data. Dey et al. (2018) propose a feature extraction method for face recognition which is applied to an LDA base classifier. Finally, references (Xu et al., 2016) and (Kanwal et al., 2021) develop linear classifiers for high-dimensional, small sample size chemometric data. Many of the aforementioned methods are variants of the classical LDA classifier which, in addition to its numerous optimality properties discussed in Chapter 2, has been shown to exhibit greater robustness than and be competitive with more sophisticated methods despite its simplicity (Lim et al., 2000; Hand, 2006).

### 1.1.3 Why asymptotic RMT?

While high-dimensional, small sample data may present a curse of dimensionality in some aspects, it also presents a blessing of dimensionality in that results from asymptotic RMT may be leveraged which would not otherwise hold in smaller dimensions.

RMT is a broad field concerned with the study of the distributions of the spectra and eigenvectors of random matrices as well as related entities such as the minimum and maximum eigenvalues. Asymptotic RMT provides asymptotic results on these quantities. What is special about these results is the growth regime; it assumes that the dimensions of the random matrix grow together. This makes

asymptotic RMT suited to high-dimensional data where the data dimensionality is on the order of its sample size. We can utilize results from asymptotic random matrix theory to derive limits of expressions involving random matrices, such as the probability of misclassification and other performance metrics. As these expressions are closed-form and deterministic, this can lead to insights into classifier performance and inspire modifications to further enhance performance and also facilitates optimization. Contrast this with a Monte Carlo simulation approach to analysis which can neither show the causal relationships and interactions between data statistics, classifier parameters, and the metric of interest, nor can it be optimized using mathematical tools (Müller and Debbah, 2016). Asymptotic RMT also allows us to derive asymptotic distributions of the classifier discriminant in the RMT regime. As we will see in Chapter 3, this can give intuition about how a classifier behaves. Finally, we can use asymptotic RMT to derive estimators for classifier performance metrics which are consistent in the RMT growth regime. These estimators are more computationally-efficient than conventional empirical procedures such as Cross-Validation (CV). Since they are consistent estimators they also do not need any additional data to be computed and can be used to tune hyperparameters using the training data without incurring selection bias which typically must be circumvented in small sample situations by using very computationally-draining techniques such as nested CV. Thus, asymptotic RMT provides an entire framework for performance analysis and design of classifiers for the current era of big data and that is what makes it so suitable for our study of linear classifiers for high-dimensional, small sample size data.

The next section goes into more detail regarding what asymptotic RMT is and what problems it can address.

## 1.2 RMT

Asymptotic RMT is our tool of choice throughout this work. To give the reader an idea of what RMT is and how it may be used, this section provides a brief

history of the field as well as a motivating example based on the sample covariance matrix.

A random matrix is a matrix whose entries are random variables. As a result, its eigenvalues (spectrum) and eigenvectors are also random variables. Random matrix theory has to do with the study of the distributions of the spectrum and eigenvectors of a random matrix and related entities such as its minimum and maximum eigenvalues (Najim and Couillet, 2018). Historically, the interest in random matrices began in statistics; the first random matrix that was studied was the Wishart matrix (Krishnapur, 2011) in 1928. This was followed by physicist Eugene Wigner's interest in random matrices as statistical models for heavy nuclei atoms in the 1950s. These were the beginnings of a mathematical theory of the spectra of random matrices (Anderson et al., 2010).

To motivate the study of the *asymptotic* spectra of random matrices, we briefly discuss an example from (Couillet and Debbah, 2011). Consider  $n$  independent and identically distributed (i.i.d.) scalar random variables,  $x_1, \dots, x_n$ , each having mean  $\mu$  and variance  $\sigma^2$ . From the strong law of large numbers, the sample mean,  $\mu_n := \frac{1}{n} \sum_{i=1}^n x_i$ , converges almost surely (a.s.) to its expectation,  $\mathbb{E}[x_1] = \mu$ , i.e., the population mean. Similarly, the sample variance,  $\sigma_n^2 = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|^2$ , converges a.s. to its expectation,  $\sigma^2$ , i.e., the population variance. In this case,  $\mu_n$  and  $\sigma_n^2$  are referred to as consistent estimators or, more specifically, as *n-consistent* estimators, as they are asymptotically and almost-surely equal to their population counterparts  $\mu$  and  $\sigma^2$ , respectively, in the limit  $n \rightarrow \infty$ .

Now, consider a sequence of  $n$  zero-mean, i.i.d. random vectors,  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{C}^p$ , each with covariance matrix  $\mathbf{\Sigma} \triangleq \mathbb{E}[\mathbf{x}_1 \mathbf{x}_1^H]$ . The sample covariance matrix  $\mathbf{R}_n := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^H$  converges a.s. to  $\mathbf{\Sigma}$  for fixed  $p$ , where the convergence is in the sense of any norm  $\|\mathbf{R}_n - \mathbf{\Sigma}\|$ .  $\mathbf{R}_n$  is, again, an  $n$ -consistent estimator of  $\mathbf{\Sigma}$ , as asymptotically, in the limit  $n \rightarrow \infty$ , for fixed  $p$ ,  $\mathbf{R}_n$  is a.s. equal to  $\mathbf{\Sigma}$  in terms of any norm. The insistence on fixed  $p$  is important; when  $n$  is large, but not much

larger than  $p$ ,  $\|\mathbf{R}_n - \mathbf{\Sigma}\|$  may be far from zero. To get some insight into how this manifests, let us take a closer look into the discrepancy in the eigenvalues between the sample covariance and the population covariance when  $p > n$ .

Say we have the same zero-mean, i.i.d. random vectors with covariance matrix  $\mathbf{\Sigma}$  and we want to estimate  $\mathbf{\Sigma}$ . The matrix  $\mathbf{R}_n$ , which converges to  $\mathbf{\Sigma}$  for fixed  $p$ , seems like the obvious choice, however, as alluded to above, in many cases this is not a good choice at all. As discussed in Section 1.1.1, there are many applications where relatively few samples are available of the data vector compared to the dimension of the vector, i.e.,  $n$  is on the order of  $p$ . To understand how  $\mathbf{R}_n$  is a bad estimator for  $\mathbf{\Sigma}$  in such cases, we rewrite the sample covariance matrix as  $\mathbf{R}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^H$  where  $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ . Specifically when  $p > n$ , the  $p \times n$  matrix  $\mathbf{X}$  has at most rank  $n$ . Since  $\text{rank}\{\mathbf{X}\} = \text{rank}\{\mathbf{X}^H\} = \text{rank}\{\mathbf{X}\mathbf{X}^H\} = \text{rank}\{\mathbf{X}^H\mathbf{X}\}$ , it follows that  $\mathbf{R}_n = \frac{1}{n}\mathbf{X}\mathbf{X}^H$  has at most rank  $n$  while  $\mathbf{\Sigma}$  has at most rank  $p$ . If  $\mathbf{\Sigma}$  is full-rank for example, then  $\mathbf{R}_n$  and  $\mathbf{\Sigma}$  differ in at least the  $p - n$  eigenvalues which we know to be zero for  $\mathbf{R}_n$ . Thus  $\mathbf{R}_n$  is not a good estimator of  $\mathbf{\Sigma}$  since they differ in their eigenvalues. In fact, the eigenvalue distribution of  $\mathbf{R}_n$  generally does not converge to the eigenvalue distribution of  $\mathbf{\Sigma}$  in any growth regime where  $p, n \rightarrow \infty$  such that  $\frac{p}{n} \rightarrow c \in (0, \infty)$ , i.e.,  $p$  and  $n$  grow commensurately, although the eigenvalue distribution of  $\mathbf{R}_n$  does generally converge. Paradoxically, the  $(i, j)$ <sup>th</sup> entry of  $\mathbf{R}_n$ ,  $(\mathbf{R}_n)_{i,j} = \frac{1}{n} \sum_{k=1}^n (\mathbf{x}_k)_i (\mathbf{x}_k)_j^*$ , does converge almost-surely to its expectation,  $\mathbb{E}[(\mathbf{x}_k)_i (\mathbf{x}_k)_j^*] = (\mathbf{\Sigma})_{i,j}$ , i.e., the  $(i, j)$ <sup>th</sup> entry of  $\mathbf{\Sigma}$ , regardless of  $p$ . This entrywise convergence always holds since  $(\mathbf{R}_n)_{i,j}$  does not depend on  $p$ .

In conclusion,  $\mathbf{R}_n$  is an  $n$ -consistent estimator, but not an  $(n, p)$ -consistent estimator for  $\mathbf{\Sigma}$ , and thus cannot be applied reliably to a finite scenario in which  $n$  is not much bigger than  $p$ . The question then arises: What distribution do the eigenvalues of  $\mathbf{R}_n$  converge to if not the distribution of its corresponding population covariance? And can we know the exact eigenvalue distribution of  $\mathbf{R}_n$  for any *finite*  $n$  and  $p$ ?

As is detailed in references (Couillet and Debbah, 2011) and (Tulino et al.,

2004), for some very basic matrices and slight variations of them, one can obtain closed-form expressions of the probability density functions of the matrix itself, its joint eigenvalues, and its marginal eigenvalues. For more complicated matrices, i.e., those having correlations and/or different variances for each entry, this is very difficult. The statistics of these matrices can be dealt with using asymptotic RMT, which may lead to deterministic limits on the spectral distributions, which can be applied to the finite-dimensional matrix case with little error. When this is not possible (because the spectrum does not converge), often one may derive Deterministic Equivalents (DEs) of functionals of certain random matrices (Couillet and Debbah, 2011). This is explained in more detail in Chapter 2. Broadly speaking, the methodology in applying asymptotic RMT is to allow both dimensions of the matrix to grow according to a growth regime where  $p, n \rightarrow \infty$  with  $\frac{p}{n} \rightarrow c \in (0, \infty)$ . We call this the *RMT growth regime*. This is in contrast to the classical growth regime where  $n \rightarrow \infty$  with  $\frac{p}{n} \rightarrow 0$ . In the RMT growth regime, both dimensions grow to infinity and are of the same order.

The reader may be wondering how knowledge of the eigenvalue distribution of a random matrix, whether exact or asymptotic, is related to the analysis of high-dimensional variants of LDA or any other classifier. In many applications, the key quantities one wishes to study can be modeled as functions of random matrices. These quantities may in turn be expressed in terms of random eigenvalues. In such cases, it becomes important to know the distribution of those eigenvalues in order to compute deterministic summaries of these quantities, such as the expectation and variance, which can yield insights into the behavior of otherwise random and obscure phenomena. There are an abundance of asymptotic results of this nature on a variety of different random matrix structures. We make use of these results in this work to analyze and estimate fundamental quantities in classification such as the probability of misclassification. Our approach for a given classifier is quite systematic and can be summarized as follows:

1. Derive the probability of misclassification of the classifier (under assump-

tions on the data distribution).

2. Derive limits of the probability of misclassification of the classifier using random matrix theory. These usually reveal insights into its behavior.
3. Derive consistent estimators in the random matrix theory growth regime, called *G-estimators*.

As we consider classification for data which is characterized by its high dimensionality, which is greater than or comparable to the number of samples of said data, the RMT growth regime is perfectly suited for its analysis.

The next chapter provides an overview of our work and contributions.

### 1.3 Overview of work and contributions

In this dissertation, we take a journey from a specific high-dimensional variant of a certain linear classifier, to a broader class of variants of that classifier, to the general class of linear classifiers, and finally to an application of the proposed technique to deep neural networks—which are not linear at all—which is possible because the last layer of a deep neural network constitutes a linear classifier. Thus, this dissertation is not arranged in chronological order, but in order of increasing generality.

The focus of Chapter 3 is a variant of the very classical and statistically motivated LDA classifier. Despite being the optimal classifier under certain Gaussian assumptions on the data, LDA performs poorly when the data dimensionality and sample size are comparable, and completely fails when the dimensionality exceeds the sample size, due to the inversion of a singular sample covariance estimate. To overcome this, many variants of LDA have been proposed over the years including alternative estimator and random projection-based variants. As these variants are intended for high-dimensional data, asymptotic RMT is the perfect tool to study them.

The variant we consider is based on randomly-projected LDA discriminants.

These classifiers solve the singularity issue by a form of dimensionality reduction using random matrices. More specifically, we consider two ensemble versions of this classifier: discriminant-averaging and vote-averaging. The main insights and findings from this work are:

- In the discriminant-averaging scheme, random projections act asymptotically as a regularization of the sample covariance estimate through the Marzetta estimator of the precision matrix (Marzetta et al., 2011) which appears in the discriminant
- The relationship between the finite version of the discriminant-averaging RP-LDA ensemble classifier and its theoretical (non-attainable) infinite version
- The optimal way to form an ensemble of RP-LDA discriminants is a form of discriminant averaging
- Projection selection within the ensemble is non-optimal
- A framework for tuning the optimal form of ensemble parameters using G-estimators

From the realization that the discriminant-averaging RP-LDA ensemble classifier is a special case of an LDA classifier modified with a general rotationally-invariant estimator in place of the precision matrix where the rotationally-invariant estimator is the Marzetta estimator, we set out in Chapter 4 to design a general LDA classifier with non-linear shrinkages that optimize the probability of misclassification. Through this work we were able to propose such a scheme as well as provide estimators for the probability of misclassification of any shrinkage scheme in conjunction with an LDA base classifier under some conditions on the shrinkage function. Previous work either considered shrinkages which optimize metrics unrelated to the classification context, linear shrinkages (as opposed to non-linear shrinkages), or shrinkages which assume underlying covariance models that are restrictive in practice.

Finally, in Chapter 5, we consider the weight vector tuning of linear classifiers. In high dimensions, tuning a classifier’s weight vector is computationally inefficient. Not only that—we do not even know where to start! This work proposes a method of tuning the multivariate weight vector through a scalar based on the decomposition of the discriminant into information and noise. This applies to any linear classifier. We show that that the method has the potential to compensate performance loss due to non-optimal native hyperparameter settings. For example, applying the weight vector tuning technique to linear SVM with non-optimal penalty brings its performance up to par with a linear SVM with optimally tuned penalty. Thus, with weight vector tuning, rather than having to tune the penalty optimally, which involves a series of optimization problems involving the large weight vector, one can simply tune the weight vector through a scalar parameter. We also demonstrate the potential for this technique in the context of deep neural networks. Pretrained deep neural networks are adapted for tasks and datasets, other than those they were trained for, by changing the configuration and retraining the weights of later layers while freezing earlier layers. This is called *transfer learning*. Usually, a small dataset is all that is available for this second phase of training, and therefore the final weights may not be trained optimally. By applying the weight vector tuning technique to the last learnable layer of a pretrained neural network that has been transfer-learned for a binary classification task, we obtain performance gains over the untuned neural network.

In the next chapter, we layout some technical background on topics that are fundamental to the rest of this dissertation, including linear classifiers, LDA, the classification setting, and some RMT concepts.

## Chapter 2

### Technical Background

In this chapter, we touch on important topics referred to in the remainder of this dissertation including linear classifiers, LDA, the classification setting, and some RMT concepts.

#### 2.1 Linear classifiers

Consider a supervised classification problem in which a test point  $\mathbf{x} \in \mathbb{R}^p$  belongs to one of two classes  $\mathcal{C}_0$  and  $\mathcal{C}_1$ . The class label,  $y \in \{0, 1\}$ , of  $\mathbf{x}$  is unknown. Our task is to build a classifier which is able to reliably predict  $y$ . A linear approach to this problem decides the class using a decision rule of the form

$$C(\mathbf{x}) = \begin{cases} 1, & \text{if } W(\mathbf{x}) > \zeta \\ 0, & \text{otherwise,} \end{cases} \quad (2.1)$$

where  $\zeta \in \mathbb{R}$  is a classifier-specific threshold and  $C(\mathbf{x}) = i$ ,  $i = 0, 1$ , indicates the class prediction, based on the *linear* discriminant

$$W(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0, \quad (2.2)$$

which is characterized by a *weight* vector,  $\mathbf{w} \in \mathbb{R}^p$ , and *bias*,  $w_0 \in \mathbb{R}$ .

#### 2.2 Classification setting

Now continuing from the previous section, assume furthermore that the two classes  $\mathcal{C}_0$  and  $\mathcal{C}_1$  constitute a Gaussian mixture having means  $\boldsymbol{\mu}_0$  and  $\boldsymbol{\mu}_1$ , respec-

tively, a common covariance  $\Sigma$ , and prior probabilities  $\pi_0$  and  $\pi_1$ , respectively, and that the test point  $\mathbf{x}$  is drawn from this Gaussian mixture model, i.e.,

$$\mathbf{x}|\mathbf{x} \in \mathcal{C}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma), \quad i = 0, 1, \quad (2.3)$$

and

$$\pi_i := P[\mathbf{x} \in \mathcal{C}_i], \quad i = 0, 1.$$

The exact testing error of a generic binary linear classifier of the form (2.1) with weight vector  $\mathbf{w}$  and intercept  $w_0$  for a given training set under these data distribution assumptions, can easily be derived as (see Lemma 1 in (Niyazi et al., 2020a))

$$\pi_0 \Phi\left(\frac{\mathbf{w}^T \boldsymbol{\mu}_0 + w_0}{\sqrt{\mathbf{w}^T \Sigma_0 \mathbf{w}}}\right) + \pi_1 \Phi\left(-\frac{\mathbf{w}^T \boldsymbol{\mu}_1 + w_0}{\sqrt{\mathbf{w}^T \Sigma_1 \mathbf{w}}}\right), \quad (2.4)$$

with  $\Sigma_0 = \Sigma_1 = \Sigma$ .

To help build our classifier, i.e., construct  $\mathbf{w}$  and  $w_0$ , we are given a set of training data,  $\mathcal{T} = \{(\mathbf{x}_j, y_j)\}_{j=1}^n$ , where  $\mathbf{x}_j$  is a data point drawn independently from the same distribution as  $\mathbf{x}$ , and  $y_j \in \{0, 1\}$  is its corresponding label. More formally,

$$\mathbf{x}_j|y_j = i \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma), \quad i = 0, 1, \quad (2.5)$$

Out of  $n$  training points,  $n_0$  points belong to class  $\mathcal{C}_0$  and  $n_1$  points to class  $\mathcal{C}_1$ . Let  $\mathbf{X}_0 \in \mathbb{R}^{p \times n_0}$  and  $\mathbf{X}_1 \in \mathbb{R}^{p \times n_1}$  be the data matrices whose columns are the individual training samples of  $\mathcal{T}$  corresponding to  $\mathcal{C}_0$  and  $\mathcal{C}_1$  respectively.

In the next section, we transition to a discussion on LDA: the Bayes classifier under the above data setting.

## 2.3 LDA

The LDA classifier is the Bayes classifier for a test point drawn from Gaussian classes with common covariance as in (2.3). In practice, the true statistics are unknown and a ‘plug-in’ version of the classifier where the sample statistics are

substituted for the true statistics is employed. We elaborate on these ideas further in the following sections followed by a description of several other optimality properties of LDA .

### 2.3.1 The Bayes classifier

The Bayes classifier, or Maximum A Posteriori (MAP) classifier is, by definition, the classifier which classifies to the class which maximizes the posterior probability of the test point. For a given set of  $K$  classes,  $\{\mathcal{C}_i\}_{i=0}^{K-1}$ , with *known* distributions, the Bayes classifier is attainable and is defined as

$$\operatorname{argmax}_{i=0,\dots,K-1} P[\mathbf{x} \in \mathcal{C}_i | \mathbf{x}] = \operatorname{argmax}_{i=0,\dots,K-1} \frac{f_i(\mathbf{x})\pi_i}{\sum_{k=1}^K f_k(\mathbf{x})\pi_l},$$

where  $f_i(\mathbf{x})$  is the class-conditional density of  $\mathbf{x}$  given  $\mathbf{x} \in \mathcal{C}_i$  and  $\pi_i$  is its corresponding prior probability.

Let  $\boldsymbol{\mu} := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ . Under the two-class data setting of Section 2.2, maximizing the posterior probability can be reduced to a decision rule on the log ratio,  $\ln \frac{P[\mathbf{x} \in \mathcal{C}_1 | \mathbf{x}]}{P[\mathbf{x} \in \mathcal{C}_0 | \mathbf{x}]}$ , which is expressed equivalently as follows, based on the assumptions of (2.3),

$$\begin{aligned} \ln \frac{P[\mathbf{x} \in \mathcal{C}_1 | \mathbf{x}]}{P[\mathbf{x} \in \mathcal{C}_0 | \mathbf{x}]} &= \ln \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} + \ln \frac{\pi_1}{\pi_0} \\ &= \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \left( \mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right) + \ln \frac{\pi_1}{\pi_0}. \end{aligned}$$

Thus the Bayes discriminant for this particular setting is defined as

$$W_{\text{Bayes}}(\mathbf{x}) := \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \left( \mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right) + \ln \frac{\pi_1}{\pi_0}, \quad (2.6)$$

with the corresponding classifier,  $C_{\text{Bayes}}(\mathbf{x})$ , taking the form (2.1) with  $\zeta = 0$ .

It is easy to see that the Bayes discriminant (2.6) for binary Gaussian classes with common covariance is linear in  $\mathbf{x}$  and therefore can be written in the form (2.2), with a weight vector of  $\mathbf{w}_{\text{Bayes}} := \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$  and an intercept of  $w_0^{\text{Bayes}} :=$

$-\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \left( \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right) + \ln \frac{\pi_1}{\pi_0}$ . As such, this classifier is known as the *linear* discriminant analysis classifier.

### 2.3.2 The LDA classifier in practice

In practice, the true statistics of the data are unknown and the LDA classifier is learned on  $\mathbf{X}_0$  and  $\mathbf{X}_1$  by computing the maximum likelihood (assuming the distribution on the training in (2.5)) estimates  $\hat{\boldsymbol{\mu}}_0$ ,  $\hat{\boldsymbol{\mu}}_1$ ,  $\hat{\boldsymbol{\Sigma}}$ ,  $\hat{\pi}_0$ , and  $\hat{\pi}_1$  of the true statistics  $\boldsymbol{\mu}_0$ ,  $\boldsymbol{\mu}_1$ ,  $\boldsymbol{\Sigma}$ , and prior probabilities  $\pi_0$  and  $\pi_1$ . These estimates are the sample means  $\hat{\boldsymbol{\mu}}_0 = \frac{1}{n_0} \mathbf{X}_0 \mathbf{1}_{n_0}$  and  $\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1} \mathbf{X}_1 \mathbf{1}_{n_1}$ , pooled sample covariance matrix  $\hat{\boldsymbol{\Sigma}} = \frac{(n_0-1)\hat{\boldsymbol{\Sigma}}_0 + (n_1-1)\hat{\boldsymbol{\Sigma}}_1}{n_0+n_1-2}$ , and the prior probability estimates  $\hat{\pi}_0 = \frac{n_0}{n}$  and  $\hat{\pi}_1 = \frac{n_1}{n}$ , respectively, where  $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n_0-1} (\mathbf{X}_0 - \hat{\boldsymbol{\mu}}_0 \mathbf{1}^T) (\mathbf{X}_0 - \hat{\boldsymbol{\mu}}_0 \mathbf{1}^T)^T$  and  $\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1-1} (\mathbf{X}_1 - \hat{\boldsymbol{\mu}}_1 \mathbf{1}^T) (\mathbf{X}_1 - \hat{\boldsymbol{\mu}}_1 \mathbf{1}^T)^T$ . In this case, the LDA discriminant rule is given by

$$W_{\text{LDA}}(\mathbf{x}) := \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \quad (2.7)$$

Its weight vector is  $\mathbf{w}_{\text{LDA}} := \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$ , where  $\hat{\boldsymbol{\mu}} := \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ , and its intercept is  $w_0^{\text{LDA}} := -\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}$ .

In an ideal world where the number of data samples  $n$  is much greater than their features  $p$ , the estimates  $\hat{\boldsymbol{\mu}}_0$ ,  $\hat{\boldsymbol{\mu}}_1$ ,  $\hat{\boldsymbol{\Sigma}}$ ,  $\hat{\pi}_0$  and  $\hat{\pi}_1$  tend to their true values and, as a result, (2.7) tends to the Bayes rule (2.6). However, when  $n$  and  $p$  are comparable, (2.7) is no longer asymptotically optimal due to estimation error in the sample means and sample covariance which render them inconsistent.

From now on, when we reference ‘LDA’, we are referring to the estimated version of the classifier,  $C_{\text{LDA}}(\mathbf{x})$ , of the form (2.1) which has (2.7) as its discriminant and  $\zeta = 0$ .

### 2.3.3 Additional optimality properties

Sections 2.3.1 and 2.3.2 walked us through how LDA with known statistics ends up being the Bayes classifier under our specific data distribution assumptions and how LDA with unknown statistics, i.e., the ‘plug-in’ version of LDA, tends

asymptotically to the Bayes classifier when  $n \rightarrow \infty$  for fixed  $p$ . In addition to these optimality properties, LDA is optimal in several other ways which require no assumptions on the distribution of the data.

Firstly, under certain conditions, the LDA weight vector,  $\mathbf{w}_{\text{LDA}}$ , is recovered as the *Fisher's linear discriminant* which optimizes the metric of class separation of data projected into one dimension proposed by Fisher in 1936 (Fisher, 1936). More specifically, Fisher's linear discriminant projects the test point in the direction

$$\mathbf{w}_{\text{Fisher}} = \underset{\mathbf{w}}{\operatorname{argmax}} \frac{\text{mean separation of projected classes}}{\text{total within-class variance of projected classes}}.$$

This weight vector coincides with  $\mathbf{w}_{\text{LDA}}$  in the binary class common covariance case, with no assumption of Gaussianity on the data.

Secondly, the LDA discriminant,  $W_{\text{LDA}}(\mathbf{x})$ , is equivalent to the optimal solution for a certain least squares formulation of the classification problem in the binary case (Vert, 2011). Letting  $y_i = -\frac{n}{n_0}$  encode training samples in  $\mathcal{C}_0$  and  $y_i = \frac{n}{n_1}$  encode training samples in  $\mathcal{C}_1$ , the optimization problem is

$$(\mathbf{w}^{\text{LS}}, w_0^{\text{LS}}) = \underset{\mathbf{w}, w_0}{\operatorname{argmin}} \sum_{i=1}^N (\mathbf{y}_i - w_0 - \mathbf{w}^T \mathbf{x})^2,$$

leading to the discriminant

$$W_{\text{LS}}(\mathbf{x}) := \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \mathbf{x} - \frac{n_0}{n} \hat{\boldsymbol{\mu}}_0 - \frac{n_1}{n} \hat{\boldsymbol{\mu}}_1 \right),$$

which coincides with (2.7) when  $\pi_0 = \pi_1$ .

With the conclusion of this section, we begin to see why LDA is such a popular choice of classifier upon which to build high-dimensional variants (as was discussed in Chapter 1). The next section briefly covers the main concepts of asymptotic RMT which can be used to analyze such variants.

## 2.4 RMT concepts

This section provides an overview of the main definitions and technical concepts of asymptotic random matrix theory. Although this dissertation mostly employs ready results from the literature so that a basic understanding of ideas such as the resolvent, DEs, and G-estimators, are sufficient to understand our derivations, we cover everything here for completeness.

### 2.4.1 Basic definitions

What follows are a list of basic definitions related to the field of asymptotic RMT, including the Empirical Spectral Distribution (ESD), Limit Spectral Distribution (LSD), and the Stieltjes transform.

#### Empirical spectral distribution

For any Hermitian matrix, one can define an empirical measure of its eigenvalues, whether the matrix is random or not. In the following, we define the empirical spectral measure and its corresponding Empirical Spectral Distribution (ESD). As we will see in the next section, the ESD for some classes of large Hermitian random matrices converges to a deterministic distribution, called the Limit Spectral Distribution (LSD) (Couillet and Debbah, 2011).

**Definition 2.4.1.** (Müller and Debbah, 2016) Consider a Hermitian matrix,  $\mathbf{X} \in \mathbb{C}^{p \times p}$ , having real eigenvalues denoted by  $\lambda_1, \dots, \lambda_p$ . The empirical measure,  $\mu_{\mathbf{X}}$ , of the eigenvalues of  $\mathbf{X}$  over the set  $\mathcal{A}$  is defined as

$$\mu_{\mathbf{X}}(\mathcal{A}) := \frac{1}{p} \sum_{i=1}^p \mathbb{1}_{\mathcal{A}}(\lambda_i),$$

where  $\mathbb{1}_{\mathcal{A}}(x)$  is the indicator function of set  $\mathcal{A}$ , i.e.,

$$\mathbb{1}_{\mathcal{A}}(x) := \begin{cases} 1, & x \in \mathcal{A}, \\ 0, & \text{otherwise.} \end{cases}$$

The corresponding empirical spectral distribution is defined in the following.

**Definition 2.4.2.** (Müller and Debbah, 2016) Consider  $\mu_{\mathbf{X}}(\mathcal{A})$  from Definition 2.4.1. The empirical spectral distribution,  $F^{\mathbf{X}}$ , of the eigenvalues of  $\mathbf{X}$  is defined as

$$F^{\mathbf{X}}(x) := \mu_{\mathbf{X}}((-\infty, x]) = \frac{1}{p} \sum_{i=1}^p \mathbb{1}_{(-\infty, x]}(\lambda_i).$$

## Limit spectral distribution

For some Hermitian  $p \times p$  random matrices denoted by  $\mathbf{X}_p$ , the corresponding ESD,  $F^{\mathbf{X}_p}$ , from Definition 2.4.2, converges to a deterministic limit  $F$  as  $p \rightarrow \infty$  (Couillet and Debbah, 2011). This limit is called the Limit Spectral Distribution (LSD). The rest of this section presents well-known LSDs including the semicircle law, the full-circle law, and the Marchenko-Pastur law.

**Theorem 2.4.1.** (Couillet and Debbah, 2011; Tulino et al., 2004) Consider a  $p \times p$  standard Wigner matrix  $\mathbf{W}_p$ , i.e., the upper-triangular entries are independent with entries  $\frac{1}{\sqrt{p}}W_{p,ij}$  such that  $\mathbb{E}[W_{p,ij}] = 0$  and  $\mathbb{E}[|W_{p,ij}|^2] = 1$ . Assume there exists  $\epsilon > 0$  such that  $\mathbb{E}[|W_{p,ij}|^{2+\epsilon}] < \infty$ ,  $\forall i, j$ . Then the empirical spectral distribution of  $\mathbf{W}_p$  converges a.s. to the semicircle law whose density is

$$f(x) = \frac{1}{2\pi} \sqrt{(4 - x^2)^+},$$

where  $(x)^+ = \max\{0, x\}$ .

If the entries are i.i.d., then the result is true without the  $(2 + \epsilon)$ <sup>th</sup>-order moment condition. Figure 2.1 obtained from (Najim and Couillet, 2018) shows a plot of the histogram of the eigenvalues of a Wigner matrix with  $p = 1500$  along with the semicircle law.

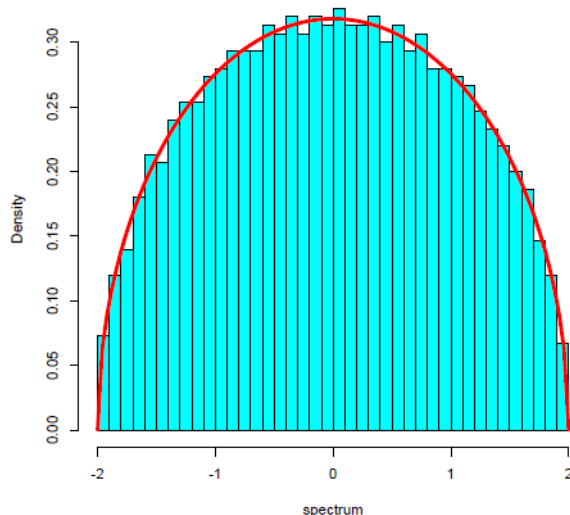


Figure 2.1: Comparison of the histogram of a Wigner matrix ( $p = 1500$ ) with the semi-circle law

The following full-circle law applies to non-Hermitian random matrices (in which case the eigenvalues are no longer real).

**Theorem 2.4.2.** (Couillet and Debbah, 2011; Tulino et al., 2004) Consider a  $p \times p$  complex matrix  $\mathbf{X}_p$  with i.i.d. entries  $\frac{1}{\sqrt{p}}X_{p,ij}$  such that  $\mathbb{E}[X_{p,ij}] = 0$ ,  $\mathbb{E}[|X_{p,ij}|^2] = 1$ , and  $\mathbb{E}[|X_{p,ij}|^6] < \infty$ . Assume also that the joint distribution of the real and imaginary parts of each entry has bounded density. The ESD of  $\mathbf{X}_p$  then converges a.s. to the full-circle law, i.e., the uniform distribution over the unit complex disk which has density  $f(x) = \frac{1}{\pi}$ , for  $\{x \in \mathbb{C} : |x| \leq 1\}$ .

Figure 2.2 obtained from (Najim and Couillet, 2018) shows a plot of the eigenvalues of a non-Hermitian matrix with  $p = 1000$  and the full-circle law.

Before moving on to the next result, we quickly define the term *Gram* matrix.

**Definition 2.4.3.** For any complex matrix  $\mathbf{X}$ , the matrix  $\mathbf{X}\mathbf{X}^H$  is defined to be the Gram matrix of  $\mathbf{X}$ .

The next result has to do with the LSD of the Gram matrix associated with a matrix whose entries are i.i.d. zero-mean and of normalized variance. This matrix is called a *Wishart* matrix and its LSD is called the Marchenko-Pastur law.

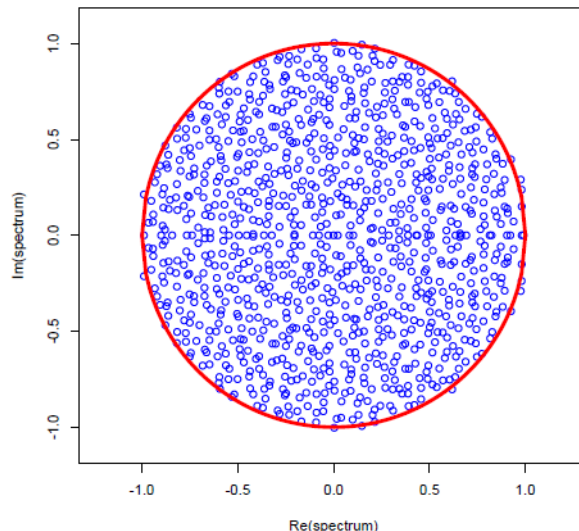


Figure 2.2: Comparison of the eigenvalue plot of a non-Hermitian matrix ( $p = 1000$ ) with the full-circle law

**Theorem 2.4.3.** (Couillet and Debbah, 2011) Consider a  $p \times n$  complex matrix,  $\mathbf{X}$ , with i.i.d. entries  $\frac{1}{\sqrt{n}}X_{p,ij}$  such that  $\mathbb{E}[X_{p,ij}] = 0$  and  $\mathbb{E}[|X_{p,ij}|^2] = 1$ . The ESD of the Gram matrix,  $\mathbf{R}_n = \mathbf{X}\mathbf{X}^H$ , of  $\mathbf{X}$  converges in distribution and a.s., as  $n, p \rightarrow \infty$  at a rate  $\frac{p}{n} \rightarrow c \in (0, \infty)$ , to the Marchenko-Pastur law whose density is

$$f_c(x) = (1 - c^{-1})^+ \delta(x) + \frac{1}{2\pi cx} \sqrt{(x - a)^+(b - x)^+}$$

where  $a = (1 - \sqrt{c})^2$ ,  $b = (1 + \sqrt{c})^2$ , and  $\delta(x)$  is the dirac delta located at zero.

Notice that  $\mathbf{R}_n$  is the sample covariance matrix of a population covariance matrix of identity. This is called the ‘null case’. The support of the Marchenko-Pastur law is

$$[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2],$$

which concentrates around 1 as  $c \rightarrow 0$ . In the limit  $c \rightarrow 0$  ( which corresponds to  $n \gg p$  in the finite regime),  $f_c(x) \rightarrow \delta(x - 1)$ , i.e., the density tends to the dirac delta located at one (Najim and Couillet, 2018). As expected, this is the spectrum of the identity matrix which is estimated perfectly by  $\mathbf{R}_n$  when  $n \rightarrow \infty$  for fixed  $p$ .

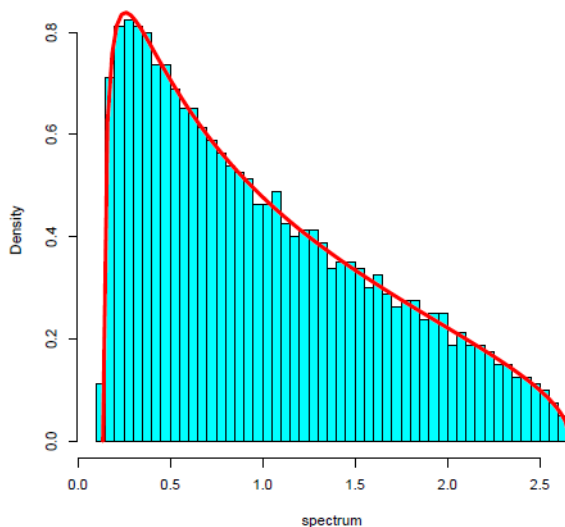


Figure 2.3: Comparison of the histogram of a Wishart matrix ( $p = 1600$ ,  $n = 4000$ ) with the Marchenko-Pastur law

Figure 2.3 obtained from (Najim and Couillet, 2018) shows a plot of the histogram of a Wishart matrix with  $p = 1600, n = 4000$  and the Marchenko-Pastur law.

For additional results on the generalizations of the Marchenko-Pastur law to the non-null case, the reader may refer to (Couillet and Debbah, 2011) and (Tulino et al., 2004).

## Stieltjes Transform

Often the limiting spectral distribution of a random matrix is obtained in the form of an invertible transform (Tulino et al., 2004). This section introduces a tool that is central to asymptotic RMT called the *Stieltjes transform*. Due to its useful properties and malleability to the methods of complex analysis, the Stieltjes transform forms the basis of the *Stieltjes transform method* for obtaining LSDs, which is outlined later in this section.

The Stieltjes transform is defined as follows.

**Definition 2.4.4.** (Müller and Debbah, 2016) For  $\mu$  a finite nonnegative measure with support,  $\text{supp}(\mu)$ , a subset of  $\mathbb{R}$ , and associated CDF,  $F_\mu(x) = \mu((-\infty, x])$ ,

the Stieltjes transform,  $m(z)$ , of  $\mu$  for  $z \in \mathbb{C} \setminus \text{supp}(\mu)$  is defined as

$$m(z) := \int_{\mathbb{R}} \frac{1}{\lambda - z} \mu(d\lambda) = \int_{\mathbb{R}} \frac{1}{\lambda - z} dF_{\mu}(\lambda).$$

For example, for  $\mu = \delta_{\lambda_0}$ , i.e, the dirac delta located at  $\lambda_0$ , the Stieltjes transform is  $m(z) = \frac{1}{\lambda_0 - z}$ ,  $z \neq \lambda_0$ . For the empirical spectral measure of Definition 2.4.1, the Stieltjes transform is  $m(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - z}$ ,  $\forall z \in \mathbb{C} \setminus \{\lambda_i\}_{i=1}^p$ . In fact, this particular Stieltjes transform can be expressed in terms of a quantity associated with every matrix called the *resolvent* which is defined in the following.

**Definition 2.4.5.** For a matrix  $\mathbf{A}_p \in \mathbb{C}^{p \times p}$ , the resolvent is defined as the matrix  $\mathbf{Q}_p(z) = (\mathbf{A}_p - z\mathbf{I}_p)^{-1}$ ,  $\forall z \in \mathbb{C}$  except for the eigenvalues of  $\mathbf{A}_p$ .

We now show how the Stieltjes transform of the empirical spectral measure of a Hermitian matrix,  $\mathbf{X}_p$ , is related to its resolvent. The resolvent,  $\mathbf{Q}_p(z)$ , of  $\mathbf{X}_p$  shares the same eigenvectors as  $\mathbf{X}_p$ . If the eigenvalues of  $\mathbf{X}_p$  are  $\lambda_1, \dots, \lambda_p$  then the eigenvalues of  $\mathbf{Q}_p(z)$  are given by  $\frac{1}{\lambda_1 - z}, \dots, \frac{1}{\lambda_p - z}$ . Matrix  $\mathbf{X}_p$  can be decomposed into  $\mathbf{X}_p = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$  and, similarly,  $\mathbf{Q}_p(z)$  can be decomposed into  $\mathbf{Q}_p(z) = \mathbf{U}(\mathbf{\Lambda} - z\mathbf{I}_p)^{-1}\mathbf{U}^H$ , where  $\mathbf{U}$  is a unitary matrix containing the shared eigenvectors and  $\mathbf{\Lambda}$  is a diagonal matrix containing the eigenvalues. Recall that the Stieltjes transform of the empirical spectral measure of  $\mathbf{X}_p$  is  $m(z) = \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - z}$ . This can be expressed

in terms of  $\mathbf{Q}_p(z)$  as

$$\begin{aligned}
m(z) &= \frac{1}{p} \sum_{i=1}^p \frac{1}{\lambda_i - z} \\
&= \frac{1}{p} \operatorname{tr} \left\{ \begin{bmatrix} \frac{1}{\lambda_1 - z} & & \\ & \ddots & \\ & & \frac{1}{\lambda_p - z} \end{bmatrix} \right\} \\
&= \frac{1}{p} \operatorname{tr}\{(\Lambda - z\mathbf{I}_p)^{-1}\} \\
&= \frac{1}{p} \operatorname{tr}\{(\Lambda - z\mathbf{I}_p)^{-1}\mathbf{U}^H\mathbf{U}\} \\
&= \frac{1}{p} \operatorname{tr}\{\mathbf{U}(\Lambda - z\mathbf{I}_p)^{-1}\mathbf{U}^H\} \\
&= \frac{1}{p} \operatorname{tr}\{(\mathbf{U}\Lambda\mathbf{U}^H - z\mathbf{U}\mathbf{U}^H)^{-1}\} \\
&= \frac{1}{p} \operatorname{tr}\{(\mathbf{X}_p - z\mathbf{I}_p)^{-1}\} \\
&= \frac{1}{p} \operatorname{tr}\{\mathbf{Q}_p(z)\}, \quad \forall z \in \mathbb{C} \setminus \{\lambda_i\}_{i=1}^p. \tag{2.8}
\end{aligned}$$

The result in (2.8) practically connects the Stieltjes transform to the spectra of Hermitian matrices; the Stieltjes transform of the empirical spectral measure of a Hermitian matrix is the normalized trace of its resolvent. Moreover, (2.8) uniquely determines the distribution function of the spectrum of  $\mathbf{X}_p$  and vice versa. The inverse Stieltjes transform exists for all  $\mu$  having a Stieltjes transform (Couillet and Debbah, 2011). More specifically, when only the Stieltjes transform  $m(z)$  of  $\mu$  is known, the following properties can be used to recover  $\mu$ :

Let  $\mu$  be a finite measure on  $\mathbb{R}$  with Stieltjes transform denoted by  $m(z)$ . From (Müller and Debbah, 2016), we have

- $\mu(\mathbb{R}) = \lim_{y \rightarrow \infty} -\mathbf{i}ym(\mathbf{i}y)$
- $\mu([a, b]) = \lim_{y \rightarrow 0^+} \frac{1}{\pi} \int_a^b \operatorname{Im}(m(x + \mathbf{i}y))dx$ , if  $a$  and  $b$  are continuity points of  $\mu$

All of the above give rise to the *Stieltjes transform method* of determining the LSD of a Hermitian matrix. If the Stieltjes transform of the empirical spectral

measure of the matrix of interest, i.e., its normalized resolvent, converges, then the limit corresponds to the Stieltjes transform of the limit spectrum of that matrix. The latter is inverted to obtain the expression for the limit spectral density. Indeed, this is the strategy behind the proof of the Marchenko-Pastur law (Couillet and Debbah, 2011). Since the distribution function,  $F$ , of a Hermitian matrix is uniquely determined by its Stieltjes transform,  $m_F$ , and vice-versa, showing convergence of the normalized resolvent of the matrix  $\mathbf{X}\mathbf{X}^H$ , as defined in Theorem 2.4.3, to the Stieltjes transform of the Marchenko-Pastur law is equivalent to showing that the empirical spectral measure of  $\mathbf{X}\mathbf{X}^H$  converges to the Marchenko-Pastur law. This is known as the *Stieltjes-Transform method*. Suffice it to say that the Stieltjes transform method does not always work as the ESD does not always converge. In such cases, the concept of Deterministic Equivalents (DEs) becomes necessary. The DE is formally defined in the following.

**Definition 2.4.6.** (Couillet and Debbah, 2011) For a sequence of Hermitian random matrices,  $\mathbf{X}_1, \mathbf{X}_2 \dots$ , with  $\mathbf{X}_p \in \mathcal{C}^{p \times p}$ , and a sequence of functionals,  $f_1, f_2, \dots$ , of  $1 \times 1, 2 \times 2 \dots$ , matrices, a Deterministic Equivalent (DE) of  $\mathbf{X}_p$  for the functional  $f_p$  is a sequence of deterministic matrices,  $\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2 \dots$ , with  $\bar{\mathbf{X}}_p \in \mathcal{C}^{p \times p}$ , such that

$$\lim_{p \rightarrow \infty} f_p(\mathbf{X}_p) - f_p(\bar{\mathbf{X}}_p) \rightarrow 0,$$

where the convergence is often in the almost-sure sense.

According to the above definition, the difference between the random functional,  $f_p(\mathbf{X}_p)$ , and the deterministic functional,  $f_p(\bar{\mathbf{X}}_p)$ , converges to zero. In this way,  $f_p(\bar{\mathbf{X}}_p)$  itself does not necessarily need to converge for it to exist. Moreover,  $f_p(\bar{\mathbf{X}}_p)$  yields an approximation of  $f_p(\bar{\mathbf{X}}_p)$  for every  $p$  which becomes increasingly more accurate with increasing  $p$  in contrast to the typical limit which summarizes an entire sequence with one statistic (Müller and Debbah, 2016). In this work, we rely heavily on existing results on DEs for various random matrix structures whose LSDs cannot be derived using the classical Stieltjes transform method.

Conventionally, the term DE is also often used to refer to the deterministic sequence  $g_1, g_2, \dots$ , where  $g_p := f_p(\bar{\mathbf{X}}_p)$  such that

$$\lim_{p \rightarrow \infty} f_p(\mathbf{X}_p) - g_p \rightarrow 0$$

in some sense (Couillet and Debbah, 2011). The DE,  $\bar{\mathbf{X}}_p$ , from Definition 2.4.6 is implicit in  $g_p$ . We adopt this convention in this work.

Finally, we briefly explain the very important concept of G-estimators. Simply put, the G-estimator is an  $(n, p)$ -consistent estimator of a functional involving a random matrix. The G-estimator is named after the mathematician Vyacheslav L. Girko who derived such estimators for many different functionals of random matrices.

The next section lists some important lemmas and DEs which are frequently referenced in our derivations.

## 2.4.2 Frequently-used lemmas and results

**Lemma 2.4.1.** (Matrix inversion lemma)(Couillet and Debbah, 2011) Let  $\mathbf{A} \in \mathbb{C}^{p \times p}$  and  $\mathbf{D} \in \mathbb{C}^{n \times n}$  be invertible and let  $\mathbf{B} \in \mathbb{C}^{p \times n}$  and  $\mathbf{C} \in \mathbb{C}^{n \times p}$ , then

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}. \end{bmatrix}$$

**Lemma 2.4.2.** (Matrix relation)(Couillet and Debbah, 2011)

$$\mathbf{I}_p - \mathbf{A}_p(\mathbf{I}_p + \mathbf{B}_p\mathbf{A}_p)^{-1}\mathbf{B}_p = (\mathbf{I}_p + \mathbf{A}_p\mathbf{B}_p)^{-1},$$

where the subscript  $p$  denotes a  $p \times p$  matrix.

**Lemma 2.4.3.** (Trace lemma)(Couillet and Debbah, 2011) Consider the sequences of matrices  $\mathbf{A}_1, \mathbf{A}_2, \dots$  with  $\mathbf{A}_p \in \mathbb{C}^{p \times p}$  with uniformly bounded spectral

norm. Let  $\mathbf{x}_1, \mathbf{x}_2, \dots$  with  $\mathbf{x}_p \in \mathbb{C}^p$  be random vectors with i.i.d. entries of zero mean, variance  $\frac{1}{p}$ , and eighth order moment of order  $O(\frac{1}{p^4})$ , independent of  $\mathbf{A}_p$ .

Then

$$\mathbf{x}_p^H \mathbf{A}_p \mathbf{x}_p - \frac{1}{p} \text{tr}(\mathbf{A}_p) \xrightarrow{\text{a.s.}} 0,$$

as  $p \rightarrow \infty$ .

**Lemma 2.4.4.** (Rank-1 perturbation lemma)(Hoydis, 2012) Let  $z \in \mathbb{C} \setminus \mathbb{R}^+$ ,  $\mathbf{A} \in \mathbb{C}^{p \times p}$  and  $\mathbf{B} \in \mathbb{C}^{p \times p}$  with  $\mathbf{B}$  a Hermitian positive semi-definite matrix and  $\mathbf{x} \in \mathbb{C}^p$ .

Then,

$$\left| \text{tr} \left( (\mathbf{B} - z\mathbf{I}_p)^{-1} - (\mathbf{B} + \mathbf{x}\mathbf{x}^H - z\mathbf{I}_p)^{-1} \right) \mathbf{A} \right| \leq \frac{\|\mathbf{A}\|}{|z|}.$$

In addition to the above lemmas, we also make frequent use of the deterministic equivalents derived in (Hachem et al., 2013), (Benaych-Georges and Couillet, 2016), and (Kammoun et al., 2019), and Lemma 8.1 of (Couillet and Debbah, 2011) throughout this dissertation.

This chapter has introduced the basic technicalities related to the work in this dissertation. The next three chapters delve into our actual contributions.

## Chapter 3

### Randomly-Projected LDA Ensembles

This chapter studies high-dimensional variants of LDA which involve random projections. More specifically, it examines two categories of RP-LDA ensembles: discriminant-averaging ensembles and vote-averaging ensembles. Through asymptotic analysis in the RMT growth regime where the problem dimensions grow at constant rates to each other for a fixed ensemble size, we determine the exact mechanism through which the ensemble size affects the classification performance. This analysis also allows us to investigate whether or not projection selection matters in an ensemble, and, ultimately, derive the optimal form of the RP-LDA ensemble. Motivated by these findings, we propose a framework for efficient tuning of the optimal RP-LDA ensemble classifier ensemble size and projection dimension based on a G-estimator of the probability of misclassification. This framework is shown to outperform the existing rule-of-thumb, as well as other methods for parameter tuning, on both real and synthetic data. An asymptotic analysis of the discriminant-averaging ensemble with infinite ensemble size in which appears the Marzetta estimator, ‘*invcov*’, of the precision matrix, is also conducted. The resulting asymptotic misclassification probability show the effect of the ensemble as a regularization of the data sample covariance matrix similar to R-LDA. Thus, the Marzetta estimator is shown to be a form of rotationally-invariant linear shrinkage asymptotically. This paves the way for the study of rotationally-invariant estimators of the precision matrix in the context of LDA in the next chapter.

Before delving into our contributions, we first provide some background concerning random projections and the literature on RP-LDA ensemble classifiers.

## 3.1 Background

### 3.1.1 Random projections

As discussed in Chapter 1, LDA performs poorly when the data dimensionality is close to the sample size, and fails altogether when it exceeds the sample size. A possible approach for adapting LDA to high-dimensional data is to reduce the data's dimensionality. Popular dimensionality reduction include procedures such as PCA, Singular Value Decomposition (SVD), the Discrete Cosine Transform (DCT) (for images), and random projection, a technique which projects the data onto a randomly selected lower-dimensional subspace. Bingham and Mannila (2001) study the effects of the aforementioned techniques on image and text data and find that, among them, random projection introduces relatively little distortion. This finding corroborates existing theory; the Johnson-Lindenstrauss lemma states that, with high probability, the distances between points in a vector space versus the points projected onto a randomly selected subspace of sufficiently high dimension are preserved. At the same time, random projection is significantly more efficient than traditional methods of dimensionality reduction. This combination of properties has the potential to yield accurate and fast classification of high-dimensional data if the data is randomly-projected before employing the LDA classifier. We therefore embark upon the study of randomly-projected LDA-based classifiers in this chapter.

To apply random projection to the data setting of Section 2.2, a matrix  $\mathbf{R} \in \mathbb{R}^{d \times p}$ , with  $d < p$ , whose entries are generated i.i.d. from a zero-mean Gaussian distribution (Durrant and Kabán, 2015), is used to multiply each data point so that it is projected onto a lower-dimensional random subspace. Projecting each training data point and re-computing the corresponding sample statistics results in the following randomly-projected statistic estimates expressed as a function of the original estimates

$$\hat{\boldsymbol{\mu}}_0^{\text{RP}} = \mathbf{R}\hat{\boldsymbol{\mu}}_0, \quad \hat{\boldsymbol{\mu}}_1^{\text{RP}} = \mathbf{R}\hat{\boldsymbol{\mu}}_1, \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}^{\text{RP}} = \mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^T,$$

corresponding to  $\hat{\boldsymbol{\mu}}_0$ ,  $\hat{\boldsymbol{\mu}}_1$ , and  $\hat{\boldsymbol{\Sigma}}$ , respectively. When the projection dimension  $d$  is chosen such that  $d \leq \text{rank}(\hat{\boldsymbol{\Sigma}})$  (noting that  $\text{rank}(\hat{\boldsymbol{\Sigma}}) \leq n-2$ ), the resulting  $\hat{\boldsymbol{\Sigma}}^{\text{RP}}$  is non-singular, since  $\mathbf{R}$  is almost surely of rank  $d$  (Durrant and Kabán, 2013). As we shall see in the next section, the randomly-projected estimators in (3.1) are plugged in to the LDA discriminant to form a single RP-LDA discriminant. Going further, one may average the discriminants obtained by projected the training data over a set of  $M$  independently-realized projections,  $\{\mathbf{R}_k\}_{k=1}^M$ , and plugging in the corresponding randomly-projected estimators. This results in an ensemble of RP-LDA discriminants which we term *the discriminant-averaging RP-LDA ensemble*. Interestingly, in the limit of the number of projections in the ensemble going to infinity, i.e.,  $M \rightarrow \infty$ , for fixed  $n$ ,  $p$ , and  $d$ , the precision matrix estimator term tends to Marzetta’s estimator of the precision matrix, ‘*invcov*’, except that we assume Gaussian random projections whereas Marzetta assumes unitary random projections drawn from Haar measure (Marzetta et al., 2011). Marzetta’s choice of distribution may be based on the fact that the original proof of the Johnson-Lindenstrauss lemma employs orthonormal random projections. This should not be an issue in practice as another proof of the Johnson-Lindenstrauss lemma exists which employs Gaussian random projections (Matoušek, 2013). In any case, when  $p$  is large, the Gaussian random projection is approximately orthonormal (Cannings and Samworth, 2017). Additionally, both distributions are invariant to orthogonal rotation. Thus we can expect the same behavior from the ensemble analyzed in this work as exhibited by *invcov*.

The other class of RP-LDA classifiers studied in this work is referred to as *vote-averaging RP-LDA ensembles*. Vote-averaging ensembles average the class prediction corresponding to each individual RP-LDA discriminant before thresholding to obtain the final prediction. Here, once again, Gaussian random projections are employed in accordance with the assumptions made in (Schclar and Rokach, 2009) and (Cannings and Samworth, 2017). We give an overview of the literature concerning these classifiers in the next section.

### 3.1.2 RP-LDA ensemble classifiers

To reiterate, the literature on RP-LDA ensemble classifiers can broadly be divided into two categories: discriminant-averaging ensembles (Durrant and Kabán, 2015; Peressutti et al., 2015) and vote-averaging ensembles (Schclar and Rokach, 2009; Cannings and Samworth, 2017). Discriminant-averaging ensembles average all RP-LDA discriminants followed by thresholding to obtain the final class prediction. Interestingly, this is equivalent to estimating the precision matrix in the LDA discriminant by the finite version of the Marzetta precision matrix estimator, *invcov*. Vote-averaging ensembles average the class prediction corresponding to each individual RP-LDA discriminant before thresholding to obtain the final prediction. The main difference between various implementations within each category is whether or not the projections are subjected to a preliminary selection process based on some criterion of expected performance within the ensemble.

Although theoretical analyses of both categories of RP-LDA ensembles exist, these studies have their limitations and raise several key questions. Durrant and Kabán (2015) derive error bounds for the basic form of the discriminant-averaging RP-LDA ensemble classifier without selection; however, the analysis is based on an abstraction wherein the ensemble size grows to infinity, revealing the converged Marzetta estimator of the precision matrix, *invcov*, within the classifier discriminant. This, in addition to a Gaussian data assumption, form the basis for the asymptotic analysis of the discriminant-averaging ensemble conducted by Niyazi et al. (2020b), where it is found that the ensemble behaves as a special case of R-LDA. This result implies that discriminant averaging can never outperform a properly-tuned R-LDA classifier on Gaussian data. Cannings and Samworth (2017) provide bounds on the error difference between the vote-averaging RP-LDA ensemble (with and without selection) and the Bayes error, but this bound is not a function of the number of projections in the ensemble. While these findings are useful, an analysis which takes into account the number of projections allows for a more accurate characterization of the practical performance of these

classifiers.

Kabán (2017, 2020) studies the performance of the finite versus the converged version of *invcov* and finds that in order to achieve a certain tolerance on the spectral norm of the difference between the two, the ensemble size must grow linearly with the data dimension. A shortcoming of this approach is that it neither provides a measure of efficacy of the finite Marzetta estimator with respect to the true measures of interest in classification, such as misclassification rate, nor does it provide practical guidelines on how to choose the number of projections. Of particular concern is the selection of an ensemble size that is small enough to maintain the computational savings provided by dimensionality reduction, yet large enough to achieve satisfactory performance.

Another gap within the literature is the lack of a thorough comparison between the discriminant averaging and vote averaging RP-LDA ensembles. An attempt at this was made in (Cannings, 2021); however, bearing in mind that the intended targets for these types of classifiers are small samples of high dimensional data, this study is very limited with regards to the dimensionality and variety of data utilized. In any case, beyond merely comparing the two types of ensembles, one would ultimately like to know the overall best way of combining any given set of RP-LDA discriminants. The next section gives an overview of contributions in this regard.

## 3.2 Contributions

The current work addresses the aforementioned issues through a comprehensive study of randomly-projected linear discriminant ensembles by asymptotic analysis under Gaussian data assumptions using RMT tools in a growth regime where the data and projection dimensions grow together. This growth regime is chosen specifically in order to more accurately represent the small-sample finite regime where the data dimensionality is greater than the number of samples, in contrast to the classical regime where the number of samples is much greater than the data

dimensionality. The analysis yields a number of insightful results stemming from the asymptotic distributions of the discriminant-averaging and vote-averaging RP-LDA ensemble classifiers and limits of their discriminant statistics and probabilities of misclassification. The main findings are:

- The class-conditional discriminant means of the discriminant-averaging RP-LDA ensemble are asymptotically identical, regardless of ensemble size.
- The asymptotic class-conditional variance of the discriminant-averaging RP-LDA ensemble is a convex combination of that of the single RP-LDA discriminant and the infinite ensemble. The asymptotic variance corresponding to the single RP-LDA discriminant is shown to be strictly greater than the asymptotic variance of the infinite ensemble.
- Each class-conditional discriminant of the discriminant-averaging RP-LDA ensemble is asymptotically Gaussian with parameters being the limits of their corresponding exact discriminant statistics.
- Each class-conditional discriminant of vote-averaging RP-LDA ensemble is asymptotically constantly-correlated Binomial with parameters being the limits of their corresponding exact probabilities of success and correlations between trials.

One of the main contributions of this study is that it shows the direct effect of the discriminant-averaging RP-LDA ensemble size on its classification performance through the asymptotic class-conditional discriminant means and variances. More specifically, since the ensemble size acts to decrease the variance of the discriminant from its maximum at a single projection to its minimum when the ensemble size grows to infinity, while maintaining a constant mean separation, the misclassification rate decreases monotonically with increasing ensemble size. Additionally, access to the single RP-LDA discriminant asymptotic distribution allows for a derivation of the asymptotically optimal way of constructing the ensemble via the Neyman-Pearson lemma and MAP rule. These results reveal

that, for Gaussian data, the optimal ensemble is linear in form, i.e., it is a form of discriminant averaging, wherein all projections are weighted equally, implying that projection selection is asymptotically sub-optimal in the context of RP-LDA ensemble classification.

The theoretical analysis in this paper also leads to several significant implications for the deployment of RP-LDA ensemble classifiers in practice, which are verified through simulations on both real and synthetic data. Firstly, our simulations suggest that there is generally no need to look beyond the basic discriminant-averaging RP-LDA ensemble classifier. Although the theoretical guarantee on which this is based assumes Gaussian data, we find that, on real data, this classifier generally performs just as well, if not better, than its immediate competitors. Secondly, as mentioned previously, classifier performance only increases with increasing ensemble size. Thus, the infinite ensemble represents the classifier's full classification potential, and finite ensemble performance may be assessed relative to the infinite ensemble. Based on this, a framework for tuning the discriminant-averaging RP-LDA ensemble size and projection dimension is proposed. An estimator of the probability of misclassification which is consistent in the RMT growth regime is derived for use in this framework. This estimator has the advantage of greater computational efficiency compared to conventional empirically-produced estimators of the test error, such as cross-validation, as well as dispensing of the need for additional data. Different variants of the tuning algorithm are implemented on real and synthetic data and compared in terms of performance and computational complexity.

The next section formally defines our problem by defining the relevant classifiers and their quantities of interest.

### 3.3 Classifier definitions

We first define the single RP-LDA discriminant followed by the two categories of ensembles of the the single RP-LDA discriminant.

### 3.3.1 The single RP-LDA classifier

Denote by  $\mathbf{R} \in \mathbb{R}^{d \times p}$  a Gaussian projection with i.i.d. entries distributed as  $\mathcal{N}(0, 1/d)$ . In order to construct a randomly-projected LDA discriminant, the training data  $\mathcal{T}$  is projected as  $\mathbf{R}\mathbf{X}_0$  and  $\mathbf{R}\mathbf{X}_1$ . The sample statistic estimates are then computed based on this projected data. It is easy to show (see, for example, Niyazi et al. (2020b)) that the resulting single randomly-projected LDA discriminant, denoted by  $W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R})$ , has the form

$$W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) = \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}, \quad (3.1)$$

where  $\hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} = \mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T)^{-1} \mathbf{R}$ . The corresponding classifier,  $C_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R})$ , then takes the form of (2.1) with a threshold of  $\zeta = 0$ .

Of interest in the analysis of this paper is the discriminant's behavior in terms of the mean separation between and variance of the distribution of projected points from each of the two classes. We denote the class-conditional means and variances of this particular discriminant by

$$m_i(1) := \mathbb{E} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i], \quad i = 0, 1 \quad (3.2)$$

and

$$\sigma^2(1) := \text{Var} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i], \quad i = 0, 1, \quad (3.3)$$

respectively, where the expectation and variance are with respect to the training data, test point, and the random projection, conditioned on the test point's class (the '1' indicates that these quantities correspond to a single random projection). As it is difficult to compute these quantities exactly, they are analyzed asymptotically in Section 3.4.1.

As mentioned previously, the randomly-projected LDA variant based on a single random projection does not perform well in practice. The two main ensemble-based schemes that have been proposed in order to improve upon the performance

of RP-LDA classifier are discriminant-averaging ensembles and vote-averaging ensembles. These are considered in the next section.

### 3.3.2 RP-LDA ensemble classifiers

#### Discriminant-averaging ensemble

A discriminant-averaging ensemble of RP-LDA discriminants averages multiple discriminants, each of which corresponds to an independently-realized random projection. In this paper, we focus on a particular discriminant-averaging ensemble which weights the contribution of each projection equally rather than according to some measure of how ‘good’ they are. In fact, we establish later in the paper that the uniformly-weighted discriminant-averaging scheme is asymptotically optimal under the data distribution assumptions detailed in Section 2.2.

Now, let us formally define the discriminant-averaging ensemble of interest. Letting  $\mathbf{R}_k$  correspond to the  $k^{\text{th}}$  projection among  $M$  random projections  $\mathbf{R}_1, \mathbf{R}_2, \dots, \mathbf{R}_M$ , the discriminant-averaging scheme which assigns equal weights to each RP-LDA discriminant is constructed as (Durrant and Kabán, 2015)

$$W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) = \frac{1}{M} \sum_{k=1}^M W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k). \quad (3.4)$$

One can imagine a scheme in which the weights of  $1/M$  in (3.4) vary for each discriminant as a function of its random projection. Furthermore, the weights may take on binary values of zero and one, thus excluding certain projections altogether. This is referred to as ‘projection selection’ in this paper. Peressutti et al. (2015) employ a kind of projection selection where selection occurs through a process of generating a projection to form a single RP-LDA discriminant, followed by subjecting the resulting classifier to a predefined threshold on training error. This is repeated until a satisfactory projection is found, and until a minimum number of satisfactory discriminants is collected. Note that there is no direct correspondence between this scheme and (3.4), as the total number of projections

is not known in advance. Note also that Peressutti et al. (2015) do not compare selection to no selection, and, in fact, select the projections based on the same data that is used to evaluate classifier performance. This results in an overestimate of the performance gain that can be attributed to selection (Cawley and Talbot, 2010). We take care to avoid this bias in this work, as is detailed in Section 3.4.2.

The discriminant (3.4) is subjected to a decision rule of the form (2.1) with  $\zeta = 0$  to obtain the final classifier,  $C_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M)$ . The recommended projection dimension setting of this classifier according to empirical observations made by Durrant and Kabán (2015) is  $d = \frac{\text{rank}\{\hat{\Sigma}\}}{2}$ . We denote the uniformly-weighted discriminant-averaging RP-LDA ensemble discriminant's class-conditional means and variances, to be analyzed in Section 3.4.1, by

$$m_i(M) := \mathbb{E} [W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i], \quad i = 0, 1 \quad (3.5)$$

and

$$\sigma^2(M) := \text{Var} [W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i], \quad i = 0, 1, \quad (3.6)$$

respectively.

Moreover, of theoretical interest in this study is the ‘infinite ensemble’ for which the number of randomly-projected LDA discriminants in the ensemble, each corresponding to an independent projection, grows to infinity. Its discriminant is defined as (Durrant and Kabán, 2015)

$$\begin{aligned} W_{M=\infty}(\mathbf{x}) &:= \lim_{M \rightarrow \infty} W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) \\ &= \hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right] \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}, \end{aligned} \quad (3.7)$$

where  $\mathbb{E}_{\mathbf{R}}[\cdot]$  is the expectation with respect to the random projection  $\mathbf{R}$ , conditioned on the training data and test point, and  $\mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right]$  is, in fact, the Marzetta estimator of the precision matrix (Durrant and Kabán, 2015; Marzetta et al., 2011). This classifier sets an upper bound on finite ensemble performance, which

may be approached by employing a very large number of projections. The work by Durrant and Kabán (2015) suggests that it suffices to use the discriminant in (3.4) with  $M = 100$  to approximate (3.7) in practice, since, according to their simulations, there is very little empirical difference between ensembles with  $M = 100$  projections versus  $M = 3000$  projections. We denote the infinite ensemble discriminant's class-conditional means and variances by

$$m_i^{M=\infty} := \mathbb{E} [W_{M=\infty}(\mathbf{x}) | \mathbf{x} \in \mathcal{C}_i], \quad i = 0, 1$$

and

$$\sigma_{M=\infty}^2 := \text{Var} [W_{M=\infty}(\mathbf{x}) | \mathbf{x} \in \mathcal{C}_i], \quad i = 0, 1,$$

respectively. The asymptotic analysis of these quantities is presented in Section 3.4.1.

## Vote-averaging ensemble

In contrast to the discriminant introduced in the previous section, which averages the RP-LDA discriminants, a vote-averaging ensemble discriminant averages the final class votes obtained by thresholding each RP-LDA discriminant. In terms of the set of  $M$  random projections  $\{\mathbf{R}_k\}_{k=1}^M$ , the uniformly-weighted vote-averaging ensemble discriminant is defined as (Cannings and Samworth, 2017)

$$W_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) = \frac{1}{M} \sum_{k=1}^M C_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k). \quad (3.8)$$

The corresponding classifier,  $C_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M)$ , takes the form of (2.1) with a threshold of  $\zeta = 0.5$ . This threshold corresponds to a majority vote.

With the aim of exploiting observed differences in classification performance among random projections, Cannings and Samworth (2017) propose a projection selection scheme on top of the basic vote-averaging RP-LDA ensemble. They generate a number  $B_1 \in \mathbb{N}$  of disjoint groups of a number  $B_2 \in \mathbb{N}$  of projections

each and select the projection from each group which yields the lowest error rate according to an error estimator of choice. The final set of projections is used to build the discriminant in (3.8) composed of a total of  $B_1$  projections. Technically, this corresponds to (3.8) with a total of  $B_1 \times B_2$  projections taking on binary weights of zeros and ones. Again, Cannings and Samworth (2017) do not compare selection to no selection in an ensemble setting. The simulations in subsequent sections of this paper look further into both the question of whether the intuitive basis for projection selection holds in an ensemble setting, and the question of how to choose the number of projections for an ensemble.

The next section presents insights into the behavior of both of these classifiers based on asymptotic analysis using RMT tools.

### 3.4 Asymptotic insights

From the asymptotic analyses of the single RP-LDA discriminant, the discriminant-averaging finite ensemble discriminant, the discriminant-averaging infinite ensemble discriminant, and the vote-averaging ensemble discriminant, this section draws several insights into the behavior of RP-LDA classifiers. The conditions under which these analyses hold are:

- (a)  $0 < \liminf \frac{p}{n} < \limsup \frac{p}{n} < \infty$
- (b)  $0 < \liminf \frac{d}{n} < \limsup \frac{d}{n} < 1$
- (c)  $0 < \liminf \frac{d}{p} < \limsup \frac{d}{p} < 1$
- (d)  $\frac{n_i}{n} \rightarrow c_i \in (0, 1), i = 0, 1$
- (e)  $\limsup_p \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 < \infty$
- (f)  $\limsup_p \|\boldsymbol{\Sigma}\|_2 < \infty$
- (g)  $\liminf_p \lambda_{\min}(\boldsymbol{\Sigma}) > 0$

The following sections detail the results of the asymptotic analysis.

### 3.4.1 Convergence of discriminant statistics and asymptotic distributions

This section presents the results concerning the convergence of the class-conditional statistics of the discriminants (3.1), (3.4), and (3.7), and their asymptotic distributions. The asymptotic distribution of (3.8) is also discussed.

Previous work (Niyazi et al., 2020b) derived deterministic equivalents for the class-conditional discriminant statistics of the discriminant-averaging RP-LDA infinite ensemble. This work extends this result by deriving the analogous results for the single RP-LDA classifier and the discriminant-averaging RP-LDA finite ensemble. Additionally, it is shown that the single RP-LDA discriminant, the discriminant-averaging RP-LDA finite ensemble discriminant, and the discriminant-averaging RP-LDA infinite ensemble discriminant, each conditioned on the class of the test point, are asymptotically Gaussian having parameters which are the deterministic equivalents of their respective (exact) statistics. This allows for comparison between the three classifiers, and thus an understanding of the effect of  $M$  on the classification. For completeness, all three sets of results are presented in what follows. The explicit expressions of the DEs are provided in the appendices.

**Theorem 3.4.1.** (Single RP-LDA discriminant asymptotic distribution)

$$1/\bar{\sigma}(1) (W_{RP-LDA}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i - \bar{m}_i(1)) \xrightarrow{d} \mathcal{N}(0, 1)$$

where  $\bar{m}_i(1)$  and  $\bar{\sigma}(1)$  are deterministic sequences of  $n$ ,  $p$ , and  $d$  such that

$$m_i(1) - \bar{m}_i(1) \xrightarrow{a.s.} 0, \quad i = 0, 1$$

and

$$\sigma^2(1) - \bar{\sigma}^2(1) \xrightarrow{a.s.} 0.$$

*Proof.* See Appendix A.1 and Appendix A.4.2. □

**Theorem 3.4.2.** (Discriminant-averaging RP-LDA finite ensemble discriminant asymptotic distribution)

$$1/\bar{\sigma}(M) (W_{disc-avg}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i - \bar{m}_i(M)) \xrightarrow{d} \mathcal{N}(0, 1)$$

where  $\bar{m}_i(M)$  and  $\bar{\sigma}(M)$  are deterministic sequences of  $n$ ,  $p$ , and  $d$  such that

$$m_i(M) - \bar{m}_i(M) \xrightarrow{a.s.} 0, \quad i = 0, 1$$

and

$$\sigma^2(M) - \bar{\sigma}^2(M) \xrightarrow{a.s.} 0.$$

*Proof.* See Appendix A.2 and Appendix A.4.2. □

**Theorem 3.4.3.** (Discriminant-averaging RP-LDA infinite ensemble discriminant asymptotic distribution)

$$1/\bar{\sigma}_{M=\infty} (W_{M=\infty}(\mathbf{x}) | \mathbf{x} \in \mathcal{C}_i - \bar{m}_i^{M=\infty}) \xrightarrow{d} \mathcal{N}(0, 1), \quad i = 0, 1$$

where  $\bar{m}_i^{M=\infty}$  and  $\bar{\sigma}_{M=\infty}^2$  are deterministic sequences of  $n$ ,  $p$ , and  $d$  such that

$$m_i^{M=\infty} - \bar{m}_i^{M=\infty} \xrightarrow{a.s.} 0, \quad i = 0, 1$$

and

$$\sigma_{M=\infty}^2 - \bar{\sigma}_{M=\infty}^2 \xrightarrow{a.s.} 0.$$

*Proof.* See Appendix A.4.3. □

Furthermore, Corollary 3.4.1 below, concerned with the relationships between the deterministic equivalents of the class-conditional discriminant statistics, follows from Theorems 3.4.1, 3.4.2, and 3.4.3:

**Corollary 3.4.1.** (Asymptotic relationships between single RP-LDA, the discriminant-averaging RP-LDA finite ensemble, and the discriminant-averaging infinite en-

semble class-conditional discriminant statistics)

$$\bar{m}_i(1) = \bar{m}_i(M) = \bar{m}_i^{M=\infty}, \quad i = 0, 1, \quad (3.9)$$

$$\bar{\sigma}^2(1) > \bar{\sigma}_{M=\infty}^2, \quad (3.10)$$

and

$$\bar{\sigma}^2(M) = \frac{1}{M}\bar{\sigma}^2(1) + \left(1 - \frac{1}{M}\right)\bar{\sigma}_{M=\infty}^2. \quad (3.11)$$

*Proof.* See Appendices A.1.1 and A.2.1 for the proof of (3.9), Appendix A.1.2 for the proof of (3.10), and Appendix A.2.2 for the proof of (3.11).  $\square$

The results of Corollary 3.4.1 along with the asymptotic distributions stated in the preceding theorems reveal that the class-conditional discriminants corresponding to single RP-LDA, the discriminant-averaging finite ensemble, and the discriminant-averaging infinite ensemble, each tend to a Gaussian distribution with common means across the discriminants. Furthermore, the single RP-LDA discriminant has a variance strictly greater than that of the discriminant-averaging infinite ensemble, while the discriminant-averaging finite ensemble is a convex combination of the two determined by coefficients  $1/M$  and  $1 - 1/M$ , respectively. Thus, as  $M$  increases, the variance of the corresponding discriminant decreases from one extreme to another, all while maintaining a constant mean separation. In light of Corollary 3.4.1, the deterministic equivalents,  $\bar{m}_i(1)$ ,  $\bar{m}_i(M)$ , and  $\bar{m}_i^{M=\infty}$ , of the class-conditional means are subsequently referred to by a common notation,  $\bar{m}_i$ ,  $i = 0, 1$ .

Figure 3.1 shows an example of the asymptotic class-conditional distributions of the discriminant-averaging RP-LDA ensemble discriminant when  $M = 10$ . The figure depicts the probability densities of the class-conditional discriminants, along with the mean separation  $\bar{m}_1 - \bar{m}_0$  and three standard deviations  $3\bar{\sigma}(M)$ . Notice that the distributions overlap. The greater the overlap, the higher the

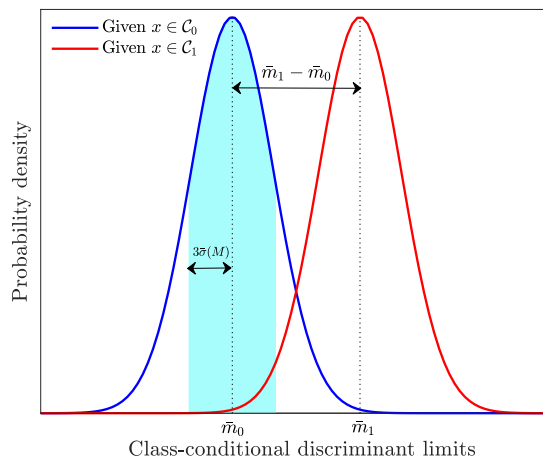


Figure 3.1: Class-conditional asymptotic distributions of the discriminant-averaging ensemble  $M = 10$ .

probability of misclassification.

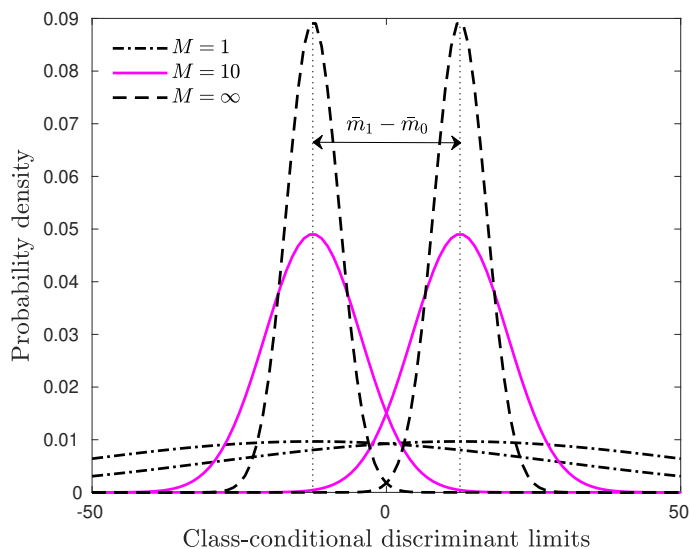


Figure 3.2: Class-conditional asymptotic distributions of the discriminant-averaging ensemble  $M = 1$ ,  $M = 10$ , and  $M = \infty$ .

Figure 3.2 shows the same distributions depicted in Figure 3.1 alongside the class-conditional asymptotic distributions of the single RP-LDA discriminant and the discriminant-averaging infinite ensemble. Consistent with Corollary 3.4.1, the mean separation between class distributions is maintained with increasing  $M$  while their variance decreases. This leads to less overlap between the distributions with increasing  $M$  and thus decreasing the probability of misclassification.

The final theorem in this section states that the asymptotic distribution of the vote-averaging RP-LDA ensemble discriminant is a correlated binomial random variable with constant correlation. To see this, we re-write the decision rule,  $C_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M)$ , as

$$C_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) = \begin{cases} 1, & \text{if } MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) > 0.5M \\ 0, & \text{otherwise,} \end{cases}$$

wherein the discriminant (3.8) as well as the decision threshold of 0.5 are scaled by  $M$ . This is clearly an equivalent classifier.

**Theorem 3.4.4.** (Vote-averaging RP-LDA ensemble discriminant asymptotic distribution)

$$P[MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i > t] - P[\mathcal{CB}(M, \bar{p}_i, \bar{\rho}_i) > t] \rightarrow 0, \forall t \in \mathbb{R}, i = 0, 1,$$

where  $M$  is the number of trials,

$$\bar{p}_i = \Phi\left(\frac{\bar{m}_i}{\sqrt{\bar{\sigma}^2(1)}}\right)$$

is the asymptotic probability of success in each trial, and

$$\bar{\rho}_i = \frac{\mathcal{I}_i - \bar{p}_i^2}{\bar{p}_i(1 - \bar{p}_i)}$$

is the asymptotic correlation between each trial, where

$$\mathcal{I}_i := \int_0^\infty \int_0^\infty \frac{1}{2\pi\bar{\sigma}^2(1)\lambda} \exp\left(-\frac{1}{2\lambda^2} \left[ \sum_{j=1}^2 \left(\frac{\alpha_j^i}{\bar{\sigma}^2(1)}\right)^2 - 2\frac{\bar{\sigma}_{M=\infty}^2 \alpha_1^i \alpha_2^i}{(\bar{\sigma}^2(1))^2} \right]\right) d\alpha_1 d\alpha_2,$$

$$\lambda^2 := 1 - \left(\frac{\bar{\sigma}_{M=\infty}^2}{\bar{\sigma}^2(1)}\right)^2, \alpha_1^i := \alpha_1 - \bar{m}_i, \text{ and } \alpha_2^i := \alpha_2 - \bar{m}_i.$$

*Proof.* See Appendix A.4.4. □

This section studies both discriminant-averaging and vote-averaging ensembles. The next section reveals that, among all RP-LDA combining schemes, the uniformly weighted discriminant-averaging RP-LDA ensemble is asymptotically optimal for Gaussian data.

### 3.4.2 Asymptotically optimal ensemble of RP-LDA discriminants

By employing individual randomly-projected LDA discriminants as observations, this section constructs the optimal ensembles in terms of the Receiver Operating Characteristic (ROC) and the probability of misclassification via the Neyman-Pearson lemma and the MAP rule, respectively. These results rely on knowledge of the asymptotic joint PDF of a collection of single RP-LDA discriminants. Letting the vector of  $M$  randomly-projected LDA discriminants be denoted by  $\mathbf{W} = [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_1), \dots, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_M)]^T$ , the asymptotic PDF of  $\mathbf{W}|\mathbf{x} \in \mathcal{C}_i$ ,  $i = 0, 1$ , is presented in the following theorem.

**Theorem 3.4.5.** (Asymptotic joint distribution of  $M$  RP-LDA discriminants)

$$P[\mathbf{W}|\mathbf{x} \in \mathcal{C}_i > t] - P[\mathcal{N}(\bar{\boldsymbol{\zeta}}_i, \bar{\boldsymbol{\Pi}}) > t] \rightarrow 0, \forall t \in \mathbb{R}, i = 0, 1,$$

where  $\bar{\boldsymbol{\zeta}}_i = \bar{m}_i \mathbf{1}_M$  and  $\bar{\boldsymbol{\Pi}} = (\bar{\sigma}^2(1) - \bar{\sigma}_{M=\infty}^2) \mathbf{I}_M + \bar{\sigma}_{M=\infty}^2 \mathbf{1}_M \mathbf{1}_M^T$ .

*Proof.* See Appendix A.4.1. □

Now, let us reconsider the classification problem in the context of hypothesis testing. Consider the null hypothesis  $\mathbf{x}$  belongs to  $\mathcal{C}_0$  and the alternative hypothesis  $\mathbf{x}$  belongs to  $\mathcal{C}_1$ . For any classifier, let  $\alpha$  be the probability of a false positive, i.e., classifying the test point to  $\mathcal{C}_1$  while it actually belongs to  $\mathcal{C}_0$ , and  $\beta$  the probability of false negative, i.e., classifying the test point to  $\mathcal{C}_0$  while it actually belongs to  $\mathcal{C}_1$ . The most powerful  $\alpha$ -level test is, by definition, the test which minimizes  $\beta$  or, equivalently, maximizes the probability of a true positive,  $1 - \beta$ , at a fixed  $\alpha$ .

Based on the asymptotic Probability Density Function (PDF) of Theorem 3.4.5, the (asymptotically) most powerful  $\alpha$ -level test is as follows.

**Theorem 3.4.6.** (Neyman-Pearson RP-LDA ensemble classifier) The (asymptotically) most powerful  $\alpha$ -level test is to classify  $\mathbf{x}$  to  $\mathcal{C}_1$  if

$$W_{disc-avg}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) > \eta$$

and to  $\mathcal{C}_0$  otherwise, where  $\eta$  is such that

$$P[W_{disc-avg}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) > \eta | \mathbf{x} \in \mathcal{C}_0] = \alpha.$$

*Proof.* According to the Neyman-Pearson lemma, for simple hypotheses, the test which rejects the null hypothesis for large values of the ratio of likelihood of observations under the alternative hypothesis to the likelihood of observations under the null hypothesis is the most powerful  $\alpha$ -level test. The likelihoods in this case are the joint PDFs of the RP-LDA discriminants under each hypothesis. Although we do not know the exact joint distributions, we do know that the discriminants are asymptotically Gaussian, as stated in Theorem 3.4.5. So, asymptotically, the likelihood ratio statistic is

$$\frac{\exp\left(-\frac{1}{2}(\mathbf{W} - \bar{\boldsymbol{\zeta}}_1)^T \bar{\boldsymbol{\Pi}}^{-1}(\mathbf{W} - \bar{\boldsymbol{\zeta}}_1)\right)}{\exp\left(-\frac{1}{2}(\mathbf{W} - \bar{\boldsymbol{\zeta}}_0)^T \bar{\boldsymbol{\Pi}}^{-1}(\mathbf{W} - \bar{\boldsymbol{\zeta}}_0)\right)} = \exp\left(\bar{m} \mathbf{1}_M^T \bar{\boldsymbol{\Pi}}^{-1} \left(\mathbf{W} - \frac{\bar{m}_0 + \bar{m}_1}{2} \mathbf{1}_M\right)\right), \quad (3.12)$$

where  $\bar{m} := \bar{m}_1 - \bar{m}_0$ . We can further simplify this by taking advantage of the special structure of  $\bar{\boldsymbol{\Pi}}^{-1}$ . Using the matrix inversion lemma (see Lemma 21 in (Müller and Debbah, 2016)) and recalling that

$$\bar{\sigma}^2(M) = \frac{1}{M} \bar{\sigma}^2(1) + \left(1 - \frac{1}{M}\right) \bar{\sigma}_{M=\infty}^2,$$

we have

$$\bar{\mathbf{\Pi}}^{-1} = \frac{1}{\bar{\sigma}^2(1) - \bar{\sigma}_{M=\infty}^2} \left[ \mathbf{I}_M - \frac{\bar{\sigma}_{M=\infty}^2}{M\bar{\sigma}^2(M)} \mathbf{1}_M \mathbf{1}_M^T \right].$$

The log of the likelihood ratio statistic in (3.12) then simplifies to

$$\frac{\bar{m}}{M\bar{\sigma}^2(M)} \sum_{k=1}^M \left[ W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k) - \frac{\bar{m}_0 + \bar{m}_1}{2} \right] = \frac{\bar{m}}{\bar{\sigma}^2(M)} W_{\text{disc-avg}}(\mathbf{x}, \{\mathbf{R}_k\}_{k=1}^M) - \frac{\bar{m}_1^2 - \bar{m}_0^2}{2\bar{\sigma}^2(M)}. \quad (3.13)$$

For  $\bar{m} > 0$ , (3.13) is an increasing function of  $W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M)$ . Thus, rejecting the null hypothesis for large values of (3.13) is equivalent to rejecting the null hypothesis for large values of  $W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M)$ . This condition can easily be verified using the definitions of  $\bar{m}_i$ ,  $i = 0, 1$ , in Appendix A.1.1. The most powerful  $\alpha$ -level test according to the Neyman-Pearson lemma is then to classify  $\mathbf{x}$  to  $\mathcal{C}_1$  if

$$W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) > \eta$$

and to  $\mathcal{C}_0$  otherwise, where  $\eta$  is such that

$$P[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) > \eta \mid \mathbf{x} \in \mathcal{C}_0] = \alpha,$$

or, equivalently (asymptotically),

$$\eta = \bar{m}_0 + \sqrt{\bar{\sigma}^2(M)} Q^{-1}(\alpha),$$

where  $Q^{-1}(\cdot)$  is the inverse Q-function. □

The above result shows that the classifier yielding the optimal ROC is in fact the uniformly-weighted discriminant-averaging RP-LDA ensemble which assigns equal weights of  $1/M$  to each of the projections  $\mathbf{R}_1, \dots, \mathbf{R}_M$ . This means that for classification purposes, in the context of an ensemble, the projections are asymptotically identical. Non-uniform weights, including binary weights, lead to asymptotically sub-optimal classification in this data setting. Note also that this

classifier is linear in the test point.

Using the asymptotic PDF of  $\mathbf{W}$ , we are also able to derive the asymptotic Bayes combination of RP-LDA discriminants which minimizes the probability of misclassification. It is presented in the following theorem.

**Theorem 3.4.7.** (MAP RP-LDA ensemble classifier) The MAP RP-LDA ensemble classifier classifies to  $\mathcal{C}_1$  when

$$\frac{\bar{m}}{\bar{\sigma}^2(M)} \left[ W_{disc-avg}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) - \frac{\bar{m}_0 + \bar{m}_1}{2} \right] + \ln \frac{\pi_1}{\pi_0} > 0, \quad (3.14)$$

and to  $\mathcal{C}_0$  otherwise.

*Proof.* The classifier which maximizes the posterior probability  $P[\mathbf{x} \in \mathcal{C}_i | \mathbf{W}]$ , minimizes the probability of misclassification (Friedman et al., 2001). Maximizing the posterior probability in the two-class scenario is equivalent to the following decision rule on the ratio of posterior probabilities

$$\frac{\pi_1 f[\mathbf{W} | \mathbf{x} \in \mathcal{C}_1]}{\pi_0 f[\mathbf{W} | \mathbf{x} \in \mathcal{C}_0]} > 0,$$

where  $f(\cdot)$  denotes a PDF. By Theorem 3.4.5, this ratio tends asymptotically to (3.14).  $\square$

Theorem 3.4.7 shows that a slight modification of (3.4) yields the asymptotically lowest probability of misclassification for an RP-LDA ensemble. Note that:

- The MAP RP-LDA ensemble classifier is also linear in  $\mathbf{x}$  and it is easy to show that  $\frac{\bar{m}_1 - \bar{m}_0}{\bar{\sigma}^2(M)} > 0$ . As a result, its ROC matches that of the discriminant-averaging RP-LDA ensemble classifier.
- This classifier corresponds to a particular operating point on the ROC of the discriminant-averaging ensemble classifier.

- When  $\pi_0 = \pi_1$ , it is easy to show that the MAP RP-LDA ensemble classifier has exactly the same decision rule as the discriminant-averaging RP-LDA ensemble classifier.

For the sake of completeness, the error analyses of the discriminant-averaging, MAP, and vote-averaging RP-LDA ensembles are detailed in Appendix A.3, wherein deterministic equivalents of the probability of misclassification are provided for each classifier.

## Demonstrations

This section showcases the results of Theorems 3.4.6 and 3.4.7 on real and synthetic data. The main idea we wish to demonstrate is that discriminant averaging is at least as good (in terms of ROC and error rate) as the discriminant-averaging-with-projection-selection, vote-averaging, and vote-averaging-with-projection-selection schemes. So that the comparison between selection and non-selection schemes is fair, we must set  $B_1 \times B_2$  in the selection schemes to  $M$ . This is because selection can be viewed as assigning weights of zeros and ones to each projection for a given set of projections, while a uniformly-weighted scheme assigns each projection in the same set of projections a weight of  $1/M$ . The total number of weighted projections in both cases must be equal; otherwise, one of the methods has the advantage of a larger initial set of projections. Thus, for the following simulations, we consider discriminant-averaging and vote-averaging ensembles with  $M = 200$  and discriminant-averaging and vote-averaging plus projection selection ensembles with  $B_1 = 50$  and  $B_2 = 4$ , so that  $B_1 \times B_2 = 200$ . The projections are selected using the resubstitution estimate on the training set.

For the synthetic data simulations, the data follows the Gaussian mixture model specified by (2.5) with

$$\boldsymbol{\mu}_0 = \frac{1}{p^{1/4}} \left[ \mathbf{1}_{\lceil \sqrt{p} \rceil}^T \quad \mathbf{0}_{p - \lceil \sqrt{p} \rceil - 2}^T \quad 2 \quad 2 \right]^T, \quad \boldsymbol{\mu}_1 = \mathbf{0}_p, \quad \text{and} \quad \boldsymbol{\Sigma} = \frac{10}{p} \mathbf{1}_p \mathbf{1}_p^T + 0.1 \mathbf{I}_p.$$

Additionally, the problem dimensions are set to  $n = 100$ ,  $p = 1000$ , and the priors to  $\pi_0 = \pi_1 = 0.5$ . The testing set consists of  $10^5$  data points. All data are generated in proportion to the prior probabilities.

Figure 3.3 shows the ROCs of each of the four classifiers. Here, the projection dimension of all four classifiers is set to  $d = 49$ , which is half the rank of the sample covariance estimate. The True Positive Rate (TPR) is plotted against the False Positive Rate (FPR). The plot shows that discriminant averaging and vote averaging perform very similarly, with discriminant averaging being slightly better. Discriminant averaging with selection and vote averaging with selection are slightly worse than the uniformly-weighted schemes. The Area Under the Curve (AUC) value corresponding to each classifier is 0.8178, 0.8100, 0.7923, and 0.7702, respectively. These results are consistent with Theorem 3.4.6, in that discriminant averaging is optimal in terms ROC and outperforms all other methods, whether they involve selection or not.

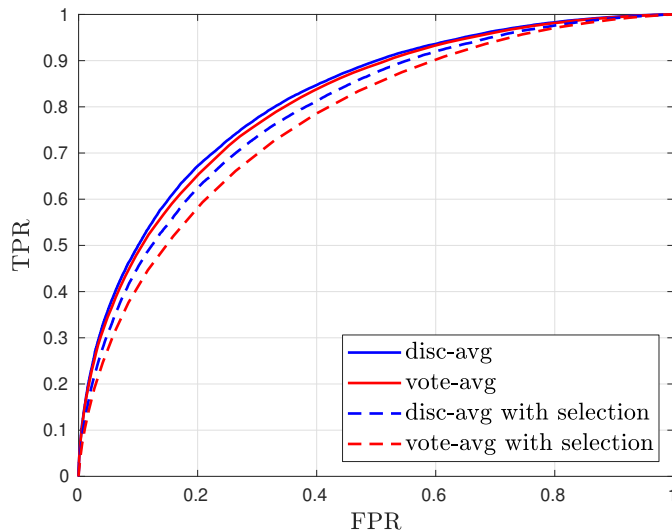


Figure 3.3: ROCs of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on Gaussian mixture model data.

Figure 3.4 plots the error rates of the four classifiers against the projection dimension  $d$ . The MAP RP-LDA ensemble is omitted as it matches discriminant averaging since the priors are assumed to be equal in this case. Again, discrimi-

nant averaging and vote averaging perform similarly, with discriminant averaging being slightly better, while the selection schemes are generally worse. Exceptions to this trend occur at relatively low values of  $d$ , more specifically at  $d = 7$  and  $d = 17$ , where discriminant-averaging with selection very slightly outperforms all other schemes. This is observed in the real data simulations in this section as well, and may be explained by the fact that our asymptotic analysis assumes that  $d$  grows commensurately with  $p$  and  $n$ , and that  $d$  here is small enough that it may be considered fixed with respect to the other dimensions in the asymptotic regime. Nevertheless, the fact that uniformly-weighted discriminant averaging generally outperforms the other classifiers is consistent with Theorem 3.4.7, since it is the MAP classifier in this equal prior scenario.

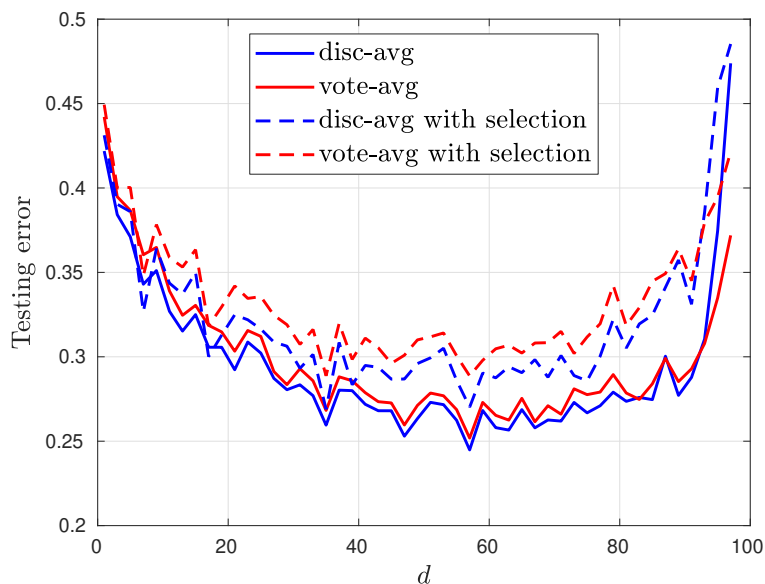


Figure 3.4: Testing error of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on Gaussian mixture model data.

For real data, we consider the colon tumor gene microarray dataset in (Alon et al., 1999), the gastrointestinal lesion colonoscopy imaging data recorded under both white light and narrow band imaging in (Mesejo et al., 2016), both the full and reduced dimension versions of the leukemia gene microarray datasets in (Golub et al., 1999), both full and reduced dimension versions of the prostate cancer gene microarray dataset in (Singh et al., 2002), and ‘aa’ and ‘ao’ phoneme

<b>Dataset</b>	$n$	$p$	<b>Proportion of majority class</b>
‘colon’	62	2000	0.65
‘gastro_WL’	76	689	0.72
‘gastro_NB’	76	689	0.72
‘leukemia_small’	72	3571	0.65
‘leukemia_big’	72	7128	0.65
‘prostate’	102	2135	0.51
‘prostate_full’	102	6032	0.51
‘phoneme_aa_ao’	100	256	0.60

Table 3.1: Datasets and their properties

pairs from the dataset (Hastie et al., 1995). These datasets are referred to as ‘colon’, ‘gastro\_WL’, ‘gastro\_NB’, ‘leukemia\_big’, ‘leukemia\_small’, ‘prostate\_full’, ‘prostate’, and ‘phoneme\_aa\_ao’ respectively. The only preprocessing done to this data consisted of removing zero-variance predictors from the gastrointestinal lesion datasets. The number of training samples, dimensionality, and proportion of data points belonging to the majority class are listed in Table 3.1. Note that ‘phoneme\_aa\_ao’ actually consists of  $n = 1717$  training samples, but to mimic a small sample situation where  $p > n$ , we randomly select a set of  $n = 100$  samples for training and utilize the remaining 1617 samples for testing. The proportion of the majority class reported in Table 3.1 for this dataset is based on the full training set. The artificially-constructed training set is sampled according to these class proportions.

Since all datasets have a relatively small number of samples, the error rates are estimated using iterated 10-fold cross validation, i.e., 10 iterations of the 10-fold cross-validation estimate are computed and averaged to obtain the final estimate, with the exception of ‘phoneme\_aa\_ao’, for which we have a test set. To avoid cross-contamination between the data used for projection selection and the data used for performance evaluation, the resubstitution error computation for projection selection is nested within the cross-validation loop (as opposed to preceding the loop) so that the selection is performed on the training folds of the cross-validation procedure at each iteration, and not on the whole training set. This is similar to the nested cross-validation procedure described in (Cawley and

Talbot, 2010).

The AUCs corresponding to the ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles applied to each of the datasets in Table 3.1 are reported in Table 3.2 along with their standard errors. The standard errors are rounded to two significant figures and each AUC is rounded up to the number of decimal places indicated by its corresponding standard error. The ‘prostate\_full’ and ‘phoneme\_aa\_ao’ datasets reflect the findings of Theorem 3.4.6 in that discriminant averaging does better than all other schemes, including discriminant averaging with selection. Interestingly, vote averaging does better than vote averaging with selection on these datasets as well. On the ‘colon’ and ‘leukemia\_small’ datasets, the AUCs corresponding to all four classifiers are virtually the same. For the datasets ‘gastro\_WL’, ‘gastro\_NB’, ‘leukemia\_big’, and ‘prostate’, we have performances which are inconsistent with Theorem 3.4.6. More specifically, on these datasets, discriminant averaging performs similarly to discriminant averaging with selection, vote averaging performs similarly to vote averaging with selection, but vote averaging performs better than discriminant averaging. This discrepancy is expected as these datasets are not necessarily Gaussian and do not necessarily meet the common covariance assumption in (2.5). Nevertheless, looking closer at the ROCs corresponding to these datasets plotted in Figures 3.5-3.7, we observe that, within the range of practical TPR and FPR, discriminant averaging performs close to, or even better than, the remaining schemes.

Figures 3.8 to 3.13 plot the iterated 10-fold CV estimates of the error rate of the discriminant-averaging, vote-averaging, and discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the real datasets listed in Table 3.1 against varying projection dimension  $d$ . As explained previously, the selection process based on the resubstitution error estimate is nested within the cross-validation in order to avoid bias due to using the same data to select the projections for and evaluate the performance of the selection classifiers. Because

Dataset	AUC			
	disc-avg	disc-avg with sel.	vote-avg	vote-avg with sel.
‘colon’	0.853 $\pm 1.5 \times 10^{-3}$	0.856 $\pm 2.1 \times 10^{-3}$	0.850 $\pm 2.9 \times 10^{-3}$	0.854 $\pm 3.2 \times 10^{-2}$
‘gastro_WL’	0.813 $\pm 5.6 \times 10^{-3}$	0.812 $\pm 5.8 \times 10^{-3}$	0.834 $\pm 6.7 \times 10^{-3}$	0.833 $\pm 5.9 \times 10^{-3}$
‘gastro_NB’	0.75 $\pm 1.1 \times 10^{-2}$	0.74 $\pm 1.2 \times 10^{-2}$	0.762 $\pm 9.8 \times 10^{-3}$	0.762 $\pm 9.9 \times 10^{-3}$
‘leukemia_small’	0.9960 $\pm 2.2 \times 10^{-4}$	0.9963 $\pm 2.3 \times 10^{-4}$	0.9960 $\pm 2.2 \times 10^{-4}$	0.9957 $\pm 4.0 \times 10^{-4}$
‘leukemia_big’	0.9911 $\pm 5.4 \times 10^{-4}$	0.9900 $\pm 5.1 \times 10^{-4}$	0.9929 $\pm 3.2 \times 10^{-4}$	0.9931 $\pm 3.7 \times 10^{-4}$
‘prostate’	0.9493 $\pm 4.9 \times 10^{-4}$	0.9518 $\pm 7.1 \times 10^{-4}$	0.958 $\pm 1.1 \times 10^{-3}$	0.961 $\pm 1.1 \times 10^{-3}$
‘prostate_full’	0.787 $\pm 4.3 \times 10^{-3}$	0.765 $\pm 5.4 \times 10^{-3}$	0.770 $\pm 5.7 \times 10^{-3}$	0.74 $\pm 8.8 \times 10^{-3}$
‘phoneme_aa_ao’	0.8478	0.8409	0.8474	0.8358

Table 3.2: AUCs corresponding to the ROCs of the discriminant-averaging, discriminant-averaging-with-selection, vote-averaging, and vote-averaging-with-selection ensembles applied to real data.

the ‘gastro\_WL’ and ‘gastro\_NB’ datasets are significantly imbalanced with 72% of the data points made up by the majority class, and error rate is not the metric of interest in such cases, these datasets are omitted in this set of simulations. The proportions of the remaining datasets are close to balanced, and so for that reason it is reasonable to assume that the uniformly-weighted discriminant-averaging ensemble performs similarly to the MAP classifier as per Theorem 3.4.7.

In all figures, it can be observed that discriminant averaging and vote averaging perform similarly, while discriminant averaging with selection and vote averaging with selection perform similarly. The selection schemes exhibit slightly lower errors than the uniformly-weighted schemes at smaller values of  $d$ . This is inconsistent with our expectation that discriminant averaging outperform all other schemes, and may be explained as before by the fact that our RMT asymptotic analysis assumes  $d$ ,  $n$ , and  $p$  to be in proportion to each other, whereas low values of  $d$  may constitute a different asymptotic regime. On the higher values of  $d$ , discriminant averaging and vote averaging outperform the selection schemes. In addition, the minimum error among all classifiers for each dataset occurs within this range. The minimum error is achieved by vote averaging at  $d = 25$  on the ‘colon’ dataset, by discriminant averaging at  $d = 27$  on the ‘leukemia\_big’ dataset, by discriminant averaging, vote averaging, and discriminant averaging with selection at  $d = 43$  on the ‘leukemia\_small’ dataset, by discriminant averaging with

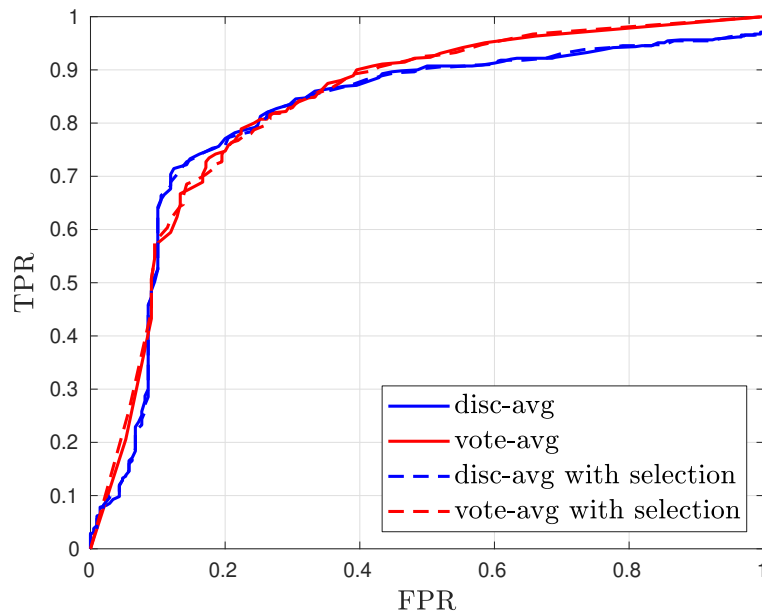


Figure 3.5: ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘gastro\_WL’ dataset.

selection at  $d = 19$  on the ‘prostate’ dataset, by vote averaging at  $d = 17$  on the ‘prostate\_full’ dataset, and by discriminant averaging at  $d = 37$ , vote averaging at  $d = 39$  and vote averaging with selection at  $d = 21$  on the ‘phoneme\_aa\_ao’ dataset. Thus, it is reasonable to conclude, that on these datasets, selection generally gives no significant advantage over uniformly-weighted schemes, and that uniformly-weighted discriminant averaging, in particular, seems to perform as well as any other scheme.

This section derives asymptotic distributions of the discriminant-averaging and vote-averaging RP-LDA ensembles. It also proves that the optimal form of RP-LDA ensemble under Gaussian data assumptions is uniformly-weighted discriminant averaging. This finding is confirmed by simulations on synthetic data, as well as on real data where it is shown that, in general, selection offers no additional performance advantage over uniformly-weighted schemes, and that discriminant averaging performs as good, if not better, than vote averaging on most datasets. Based on these findings, the next section studies the uniformly-weighted discriminant-averaging RP-LDA ensemble classifier from a practical perspective,

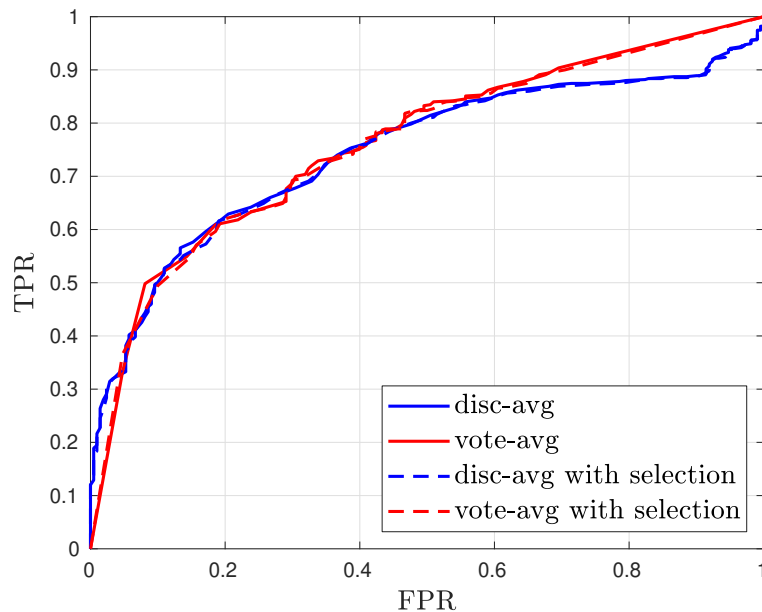


Figure 3.6: ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘gastro\_NB’ dataset.

mainly choosing the number of projections  $M$  and tuning the projection dimension  $d$  through the use of G-estimators.

### 3.5 Turning theory into practice

In this section, we focus on practical implications of the analysis of the previous section in conjunction with G-estimators to propose a working framework for RP-LDA ensemble classification. The main lessons to take from the previous section are:

1. The optimal ensemble under Gaussian data assumptions is a linear function of the RP-LDA discriminants, i.e., it is a form of discriminant averaging, as opposed to a non-linear scheme like vote averaging. As demonstrated in the previous section, both schemes perform very similarly on real data. Thus, there is no need to look beyond linear schemes.
2. Derivations under Gaussian data assumptions show that it is the number of projections which is critical for the classification performance of the

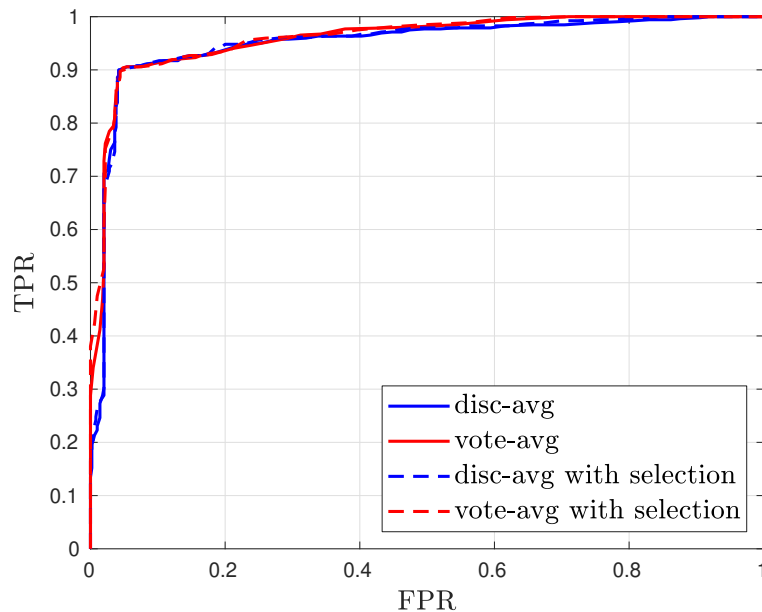


Figure 3.7: ROCs of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘prostate’ dataset.

discriminant-averaging RP-LDA ensemble, not the projections themselves. In fact, as evidenced by the previous section, projection selection under these assumptions may result in a performance loss. Furthermore, projection selection adds an extra cost in the form of computing error estimators for each single RP-LDA ensemble classifier corresponding to a member of the set of projections in order to implement the selection process.

Based on these findings, this section proposes methods for the practical implementation of the uniformly-weighted discriminant-averaging RP-LDA ensemble. We first present G-estimators of the most common classification metrics of this classifier. We then propose and demonstrate a method for tuning the number of projections  $M$  and projection dimension  $d$  on real and synthetic data.

### 3.5.1 G-estimators

As explained in Chapter 2, a G-estimator of a quantity is an estimator of that quantity which is consistent in the RMT regime. This section provides G-estimators of the class-conditional discriminant statistics of the uniformly-weighted

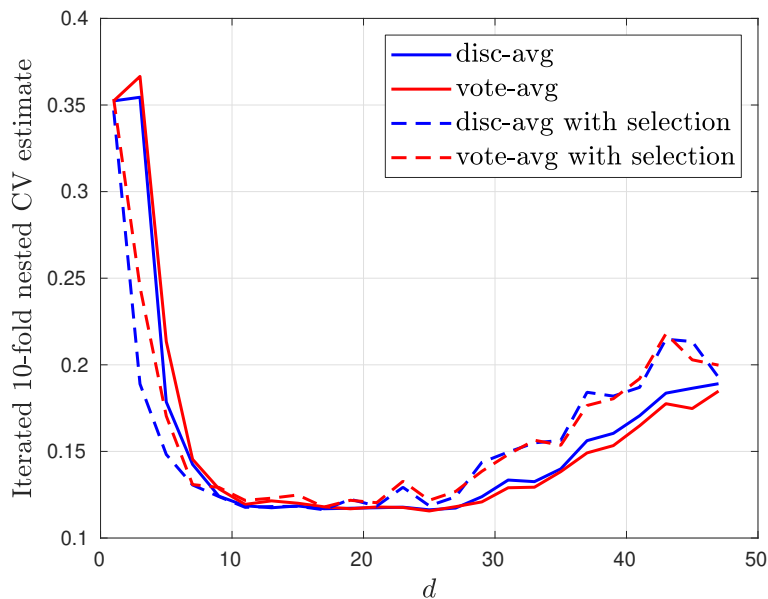


Figure 3.8: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘colon’ dataset.

discriminant-averaging RP-LDA ensemble classifier, from which G-estimators of metrics such as the TPR, FPR, True Negative Rate (TNR), False Negative Rate (FNR), probability of misclassification, Positive Predictive Value (PPV), and Negative Predictive Value (NPV) are constructed. This is detailed in what follows.

The main building blocks of the G-estimators of interest are G-estimators  $\hat{m}_i$ ,  $i = 0, 1$ ,  $\hat{\sigma}^2(1)$ , and  $\hat{\sigma}_{M=\infty}^2$  such that

$$\hat{m}_i - \bar{m}_i \xrightarrow{a.s.} 0, \quad i = 0, 1,$$

which implies  $\hat{m}_i \asymp m_i(1)$ ,  $\hat{m}_i \asymp m_i(M)$ , and  $\hat{m}_i \asymp m_i^{M=\infty}$ , i.e.,  $\hat{m}_i$  is a G-estimator of all three classifier class-conditional means,

$$\hat{\sigma}^2(1) - \sigma^2(1) \xrightarrow{a.s.} 0,$$

and

$$\hat{\sigma}_{M=\infty}^2 - \sigma_{M=\infty}^2 \xrightarrow{a.s.} 0.$$

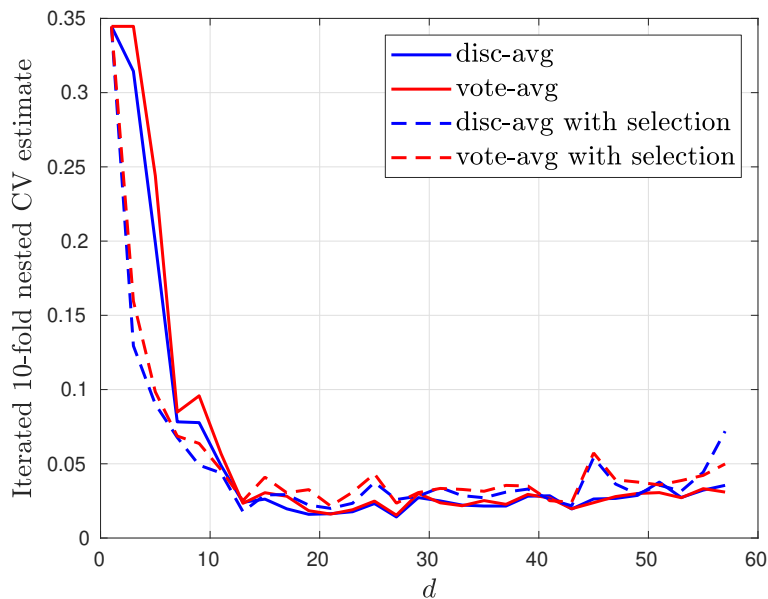


Figure 3.9: Iterated 10-fold CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘leukemia.big’ dataset.

Theorem 3.5.1 presents the explicit expressions of these G-estimators. Note that  $\hat{m}_i$  and  $\hat{\sigma}_{M=\infty}^2$  were derived in (Niyazi et al., 2020b). They are restated here for completeness.

**Theorem 3.5.1.** (G-estimators of the class-conditional discriminant statistics)

$$\hat{m}_i = (-1)^{i+1} \left[ \frac{1}{2} \hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}} - \frac{\frac{1}{n_i-1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \right] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}, \quad i = 0, 1,$$

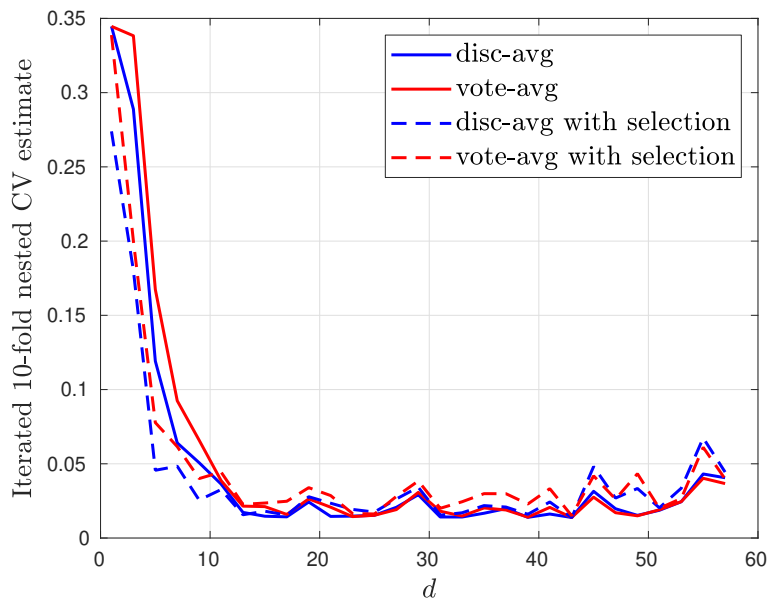


Figure 3.10: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘leukemia\_small’ dataset.

$$\hat{\sigma}^2(1) = \left( \frac{1}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \right)^2 \hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}}$$

$$+ \frac{1}{\hat{\nu}^2} \frac{\left( \frac{1}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \right)^2 \frac{1}{p} \text{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}}{1 - \frac{1}{d} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-2} \hat{\boldsymbol{\mu}},$$

and

$$\hat{\sigma}_{M=\infty}^2 = \left( 1 + \frac{\frac{1}{n-2} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \right)^2 \hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}},$$

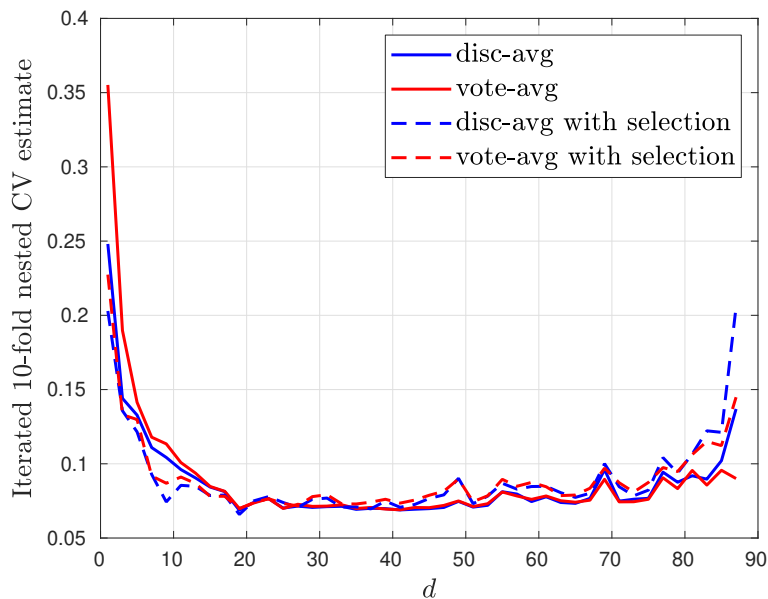


Figure 3.11: Iterated 10-fold nested CV estimate of the error rate of discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘prostate’ dataset.

where  $\hat{\nu}$  is such that

$$1 - \frac{1}{d} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\} = 0.$$

*Proof.* See Appendix A.5.1. □

In addition, it can be shown that the G-estimator  $\hat{\sigma}^2(M)$  of  $\sigma^2(M)$  such that

$$\hat{\sigma}^2(M) - \sigma^2(M) \xrightarrow{a.s.} 0.$$

is simply

$$\hat{\sigma}^2(M) = \frac{1}{M} \hat{\sigma}^2(1) + \left( 1 - \frac{1}{M} \right) \hat{\sigma}_{M=\infty}^2.$$

The next theorem presents the G-estimators of some common binary classification metrics of the uniformly-weighted discriminant-averaging RP-LDA ensemble in terms of the preceding G-estimators. Note that we take  $\mathcal{C}_0$  to be the negative class and  $\mathcal{C}_1$  to be the positive class.

**Theorem 3.5.2.** (G-estimators of some common classification metrics of the

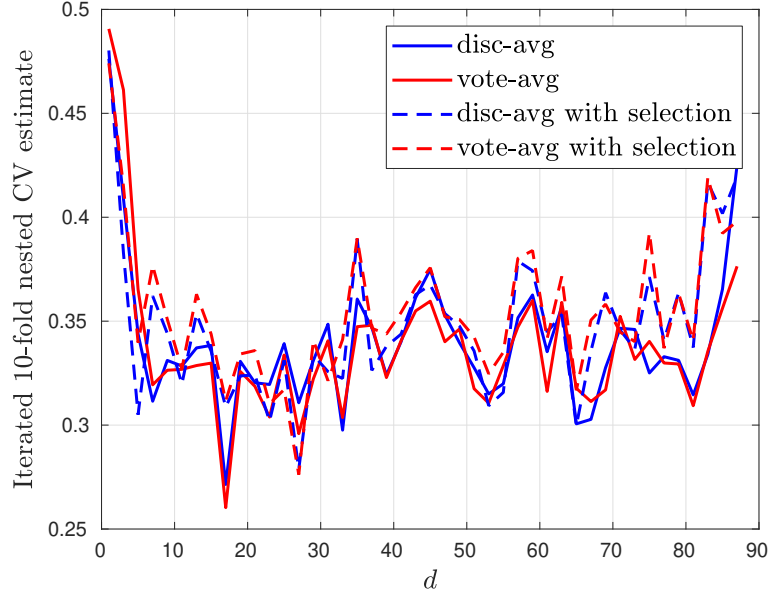


Figure 3.12: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘prostate\_full’ dataset.

discriminant-averaging RP-LDA Ensemble)

- TPR:  $\hat{\text{TPR}} = \Phi\left(\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)$
- TNR:  $\hat{\text{TNR}} = \Phi\left(-\frac{\hat{m}_0}{\sqrt{\hat{\sigma}^2(M)}}\right)$
- FPR:  $\hat{\text{FPR}} = \Phi\left(\frac{\hat{m}_0}{\sqrt{\hat{\sigma}^2(M)}}\right)$
- FNR:  $\hat{\text{FNR}} = \Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)$
- Probability of misclassification:  $\hat{\varepsilon} = \hat{\pi}_0 \hat{\text{FPR}} + \hat{\pi}_1 \hat{\text{FNR}}$
- PPV:  $\hat{\text{PPV}} = \frac{\hat{\pi}_1 \hat{\text{TPR}}}{\hat{\pi}_0 \hat{\text{FPR}} + \hat{\pi}_1 \hat{\text{TPR}}}$
- NPV:  $\hat{\text{NPV}} = \frac{\hat{\pi}_0 \hat{\text{TNR}}}{\hat{\pi}_0 \hat{\text{TNR}} + \hat{\pi}_1 \hat{\text{FNR}}}$

*Proof.* See Appendix A.5.2. □

In the next section, we propose a general procedure for tuning the parameters of the discriminant-averaging RP-LDA ensemble. We also demonstrate how G-estimators may be made use of in this context.

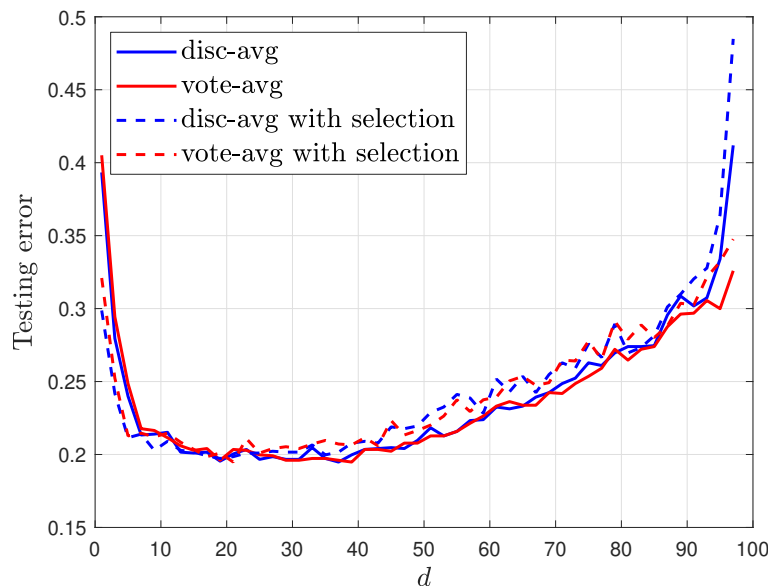


Figure 3.13: Iterated 10-fold nested CV estimate of the error rate of the discriminant-averaging, vote-averaging, discriminant-averaging-with-selection, and vote-averaging-with-selection ensembles on the ‘phoneme\_aa\_ao’ dataset.

### 3.5.2 Tuning the discriminant-averaging RP-LDA ensemble parameters

This section maps out a procedure for tuning the number of projections  $M$  and projection dimension  $d$  of the uniformly-weighted discriminant-averaging RP-LDA ensemble classifier. To begin with, note that, as shown in Section 3.4, for a given  $d$ , performance improves with increasing  $M$  so that the upper bound on the performance of the finite ensemble is the performance of the infinite ensemble. In other words, the ratio of infinite ensemble error to finite ensemble error is always less than or equal to one.

As  $M$  is constrained by computational efficiency, one may specify the trade-off between performance and computational efficiency as a fraction of the infinite ensemble performance, denoted by  $\psi$ . More specifically,  $\psi = \frac{\text{infinite ensemble error}}{\text{finite ensemble error}}$ . This idea is illustrated in Figure 3.14, which shows this ratio approaching 1 with increasing  $M$ . As alluded to earlier, the infinite ensemble cannot be realized practically, but must be approximated by a large number of projections. Figure 3.14 uses 3000 projections to approximate the infinite ensemble. As indicated on the

figure, a performance of  $\psi = 0.98$  is achieved at  $M = 112$ . This is a significant computational savings as compared to utilizing the full set of 3000 projections which approximate the infinite ensemble. Based on this, we propose the following experimental approach to tuning  $M$  and  $d$ , which uses 5000 projections to approximate the infinite ensemble. It is important to realize that this procedure is part of the classifier training, and should be applied to the training set.

1. Tune  $d$  for a ensemble with  $M = 5000$ . This approximates the optimal projection dimension for an infinite ensemble. Compute the corresponding error, which serves as the infinite ensemble performance benchmark against which we measure  $\psi$ .
2. Now starting at a small  $M$ , compute the error and the resulting ratio of infinite ensemble error (from step 1) to this finite ensemble error. Check if  $\psi$  is satisfied. If not, increment  $M$ . Repeat in this manner until  $\psi$  is satisfied. The value of  $M$  at which  $\psi$  is satisfied is the final setting of  $M$  for the finite ensemble which achieves at least  $\psi$  level of performance relative to the infinite ensemble.

While it is possible to add a third step to this procedure in which  $d$  is further tuned for the finite ensemble obtained at step 2, we find that this can result in performance loss in practice, probably due to overfitting to the training data.

Algorithm 1 presents the previously outlined procedure in more detail. Here,  $R(M, d)$  is any training-data-based estimate of the probability of misclassification of a discriminant-averaging RP-LDA ensemble composed of  $M$  projections having projection dimension  $d$ . Of course, Algorithm 1, especially lines 1-3, may be very computationally intensive depending on the choice of error estimator  $R(M, d)$ . The G-estimator of the infinite ensemble error derived in Niyazi et al. (2020b) and the G-estimator for the finite ensemble error derived in this work (see Section 3.5.1, Theorem 3.5.2) may be of utility in this context. We propose using the former for the calculation in line 2, and the latter for the calculations in lines 7 and 10. For further savings, the  $M$  tuning procedure in lines 6-12 may be

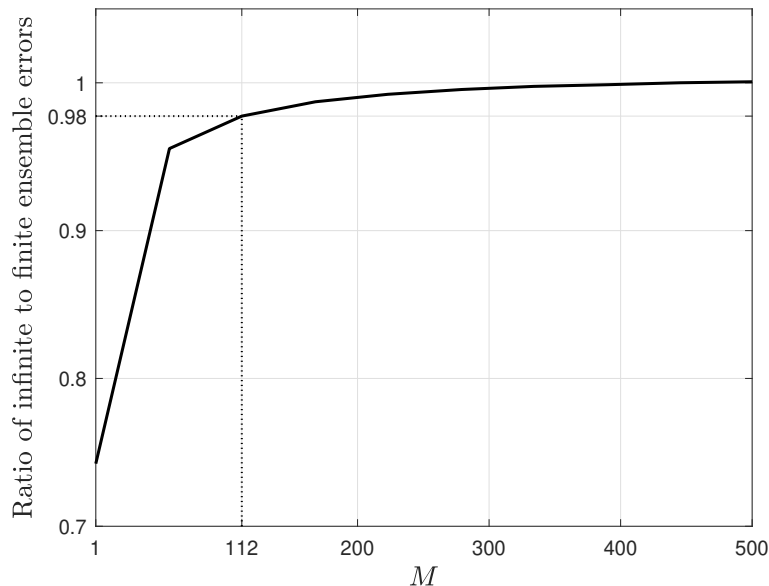


Figure 3.14: Ratio of infinite to finite discriminant-averaging RP-LDA ensemble classifier error on Gaussian mixture model data.

bypassed by using an approximation of the probability of misclassification based on G-estimators which gives

$$M \approx \text{ceil} \left( \frac{(\hat{\sigma}^2(1) - \hat{\sigma}_{M=\infty}^2) W_0 \left( \psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \exp \left( \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \right) \right)}{\hat{m}_1^2 - \hat{\sigma}_{M=\infty}^2 W_0 \left( \psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \exp \left( \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \right) \right)} \right),$$

where  $W_0(\cdot)$  is the principal branch of the Lambert W function. This approximation is valid when the class priors are **equal**. It is derived in Appendix A.5.3. In what follows, we refer to this approximation as ‘the heuristic’. All test errors are rounded to three decimal places.

We now report the errors achieved by tuning the discriminant-averaging ensemble classifier by Algorithm 1 on both real and synthetic data. We first consider synthetic data generated from the Gaussian mixture model specified at the beginning of Section 3.4.2. The testing error is evaluated on a testing set consisting of  $10^5$  data points from each class. Table 3.3 presents the testing errors and parameter settings of the infinite discriminant-averaging ensemble, R-LDA, and various tunings of the finite discriminant-averaging ensemble. The R-LDA precision matrix estimator takes the form  $(\hat{\Sigma} + \gamma \mathbf{I}_p)^{-1}$ , where  $\gamma$  is a positive scalar.

**Algorithm 1** Tuning the discriminant-averaging ensemble parameters  $M$  and  $d$ 


---

**Require:**  $\psi < 1$

- 1: **for**  $d' = 1 : \text{rank}\{\hat{\Sigma}\}$  **do**
- 2:     Compute  $R(5000, d')$  ▷ Set  $M = 5000$  to approximate an infinite ensemble
- 3: **end for**
- 4:  $d \leftarrow \text{argmin}_{d'} R(5000, d')$  ▷ Tune  $d$  for infinite ensemble
- 5:  $R_{M=\infty} \leftarrow R(5000, d)$  ▷ Set minimum infinite ensemble error estimate
- 6:  $M' \leftarrow 100$
- 7: Compute  $R(M', d)$
- 8: **while**  $\frac{R_{M=\infty}}{R(M', d)} < \psi$  **do** ▷ Set  $M$  so that  $\psi$  is satisfied
- 9:      $M' \leftarrow M' + 100$
- 10:     Compute  $R(M', d)$
- 11: **end while**
- 12:  $M \leftarrow M'$

---

The parameter setting of the infinite ensemble is determined by setting  $M = 5000$  and tuning  $d$  optimally according to the iterated 10-fold CV estimate of error on the training set. Similarly, the parameter setting of R-LDA is determined by tuning  $\gamma$  optimally according to the iterated 10-fold CV estimate of error on the training set. The parameters settings of the finite ensemble are determined by different variants of Algorithm 1 at  $\psi = 0.95$ . The first sub-row under the finite ensemble utilizes the iterated 10-fold CV on the training set as the error estimator  $R(M, d)$ . The second sub-row uses the G-estimator for the infinite and finite ensembles on the training set in place of all error estimators in Algorithm 1. The third sub-row uses the heuristic described previously to compute  $M$  directly and the G-estimators for the infinite ensemble in the second step. Finally, the fourth sub-row uses Durrant’s rule-of-thumb (see Section 3.3.2) to set  $M$  and  $d$  without any need for error estimation. To reduce fluctuation due to projections, the ‘Test error’ values reported in the table are testing errors averaged over 500 sets of projections, except for the first row which is averaged over 5 sets, since each set consists of 5000 projections.

Table 3.3 shows that R-LDA achieves the lowest testing error of 0.214 followed by the finite ensemble tuned by heuristic at 0.218. As the computational complexity of R-LDA is  $\mathcal{O}(p^3)$ , while that of the discriminant-averaging RP-LDA

Classifier		Test error	Parameters
disc-avg infinite ensemble tuned by CV		0.223	$d = 31$
R-LDA		0.214	$\gamma = 1.03$
disc-avg finite ensemble tuned by:	CV	0.241	$M = 200, d = 31$
	G-estimators	0.221	$M = 400, d = 66$
	G-estimators + Heuristic	0.218	$M = 552, d = 66$
	Durrant's rule-of-thumb	0.245	$M = 100, d = 49$

Table 3.3: Table of average testing errors and parameter settings of the infinite discriminant-averaging ensemble classifier, where  $d$  is tuned based on cross-validation, and the finite discriminant-averaging ensemble classifier where  $M$  and  $d$  are tuned by Algorithm 1 based on cross-validation, G-estimators, and the heuristic on Gaussian mixture model data. Here  $\psi = 0.95$ . For comparison, the Bayes error is 0.0401.

ensemble is  $O(M(np + d^3))$  (Durrant and Kabán, 2015), using a large  $M$  might offset the computational savings offered by the ensemble, in which case one might be better off using R-LDA. In this case, the complexity of R-LDA is on the order of  $10^9$  operations, while that of the ensemble tuned by the heuristic is on the order of  $10^8$  operations. The heuristic also has the advantage of minimal tuning at training time. The worst performance belongs to Durrant's rule-of-thumb which yields an error of 0.245. Its complexity at execution is, however, on the order of  $10^7$  and, moreover, it involves no tuning procedure. The finite ensemble tuned by cross-validation also has a relatively high test error of 0.241 at a corresponding complexity at execution on the order of  $10^7$ , although the training procedure is much more involved. The infinite ensemble tuned by cross-validation and the finite ensemble tuned by G-estimators yield test errors of 0.223 and 0.221, respectively, at a common complexity at execution on the order of  $10^8$ , and so have higher errors in addition to offering no performance advantage over the finite ensemble tuned by the heuristic.

For real data, we consider the 'phoneme\_aa\_ao' dataset. Again, we use 5000 projections to approximate the infinite ensemble. Table 3.4 can be interpreted exactly as Table 3.3. Here  $\psi = 0.99$  and, as in the synthetic data simulation,

Classifier		Test error	Parameters
disc-avg infinite ensemble tuned by CV		0.206	$d = 11$
R-LDA		0.210	$\gamma = 4.76$
disc-avg finite ensemble tuned by:	CV	0.206	$M = 2500, d = 11$
	G-estimators	0.199	$M = 100, d = 31$
	G-estimators + Heuristic	0.199	$M = 82, d = 31$
	Durrant's rule-of-thumb	0.213	$M = 100, d = 50$

Table 3.4: Table of average testing errors and parameter settings of the infinite discriminant-averaging ensemble classifier, where  $d$  is tuned based on cross-validation, and the finite discriminant-averaging ensemble classifier where  $M$  and  $d$  are tuned by Algorithm 1 based on cross-validation, G-estimators, and the heuristic on the ‘phoneme\_aa\_ao’ dataset. Here  $\psi = 0.99$ .

the reported testing errors are averaged over 500 trials except for the first row where the testing error is averaged over 5 trials. All errors are rounded to three significant figures.

Table 3.4 shows that, although the data is slightly imbalanced, the two strategies of tuning a finite ensemble using the G-estimators and the heuristic both achieve the lowest testing error of 0.199. Both have complexities at execution on the order of  $10^6$ . The fact that these finite ensembles perform better than the infinite ensemble tuned by cross-validation can be explained by the better choice of  $d$  obtained by the G-estimators ( $d = 31$ ) as compared to cross-validation ( $d = 11$ ). This is confirmed by computing the testing error of an ensemble with  $M = 5000$  and  $d = 31$  which turns out to be 0.196. Durrant’s rule-of-thumb has the worst performance on this data, yielding an error of 0.213 at a complexity at execution time, on the order of  $10^7$  operations. The infinite ensemble, R-LDA and the finite ensemble tuned by cross-validation have complexities at execution on the order of  $10^8$ ,  $10^7$ , and  $10^7$ , respectively, and so have higher errors than the G-estimator and heuristic schemes while offering no performance advantage.

The MATLAB code for the tuning framework based on the G-estimator of error and the heuristic can be found at [https://github.com/niyazil/DA-RP-ensemble\\_](https://github.com/niyazil/DA-RP-ensemble_)

tuning.

The next section concludes this chapter with an additional finding regarding the infinite discriminant-averaging RP-LDA ensemble which relates it to R-LDA.

### 3.6 The discriminant-averaging RP-LDA infinite ensemble as a special case of R-LDA

In this section, we conduct a deeper investigation of the effect of Marzetta's precision matrix estimator, *invcov*, within the context of the LDA classifier. In other words, we study the effect of the infinite ensemble term,  $\mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right]$ , on the performance of the infinite discriminant-averaging RP-LDA ensemble classifier defined at the end of Section 3.3.2. We find that this classifier asymptotically behaves (in terms of its probability of misclassification) as a special case of R-LDA with a coarser tuning of the regularization parameter.

To do this, we look at the form of the probability of misclassification of the infinite discriminant-averaging RP-LDA ensemble. Using (2.4), the exact testing error of the classifier (3.7) for a given training set,  $\mathcal{T}$ , is

$$\begin{aligned} \varepsilon_{M=\infty}(\mathcal{T}) := & \pi_0 \Phi \left( \frac{\hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right] (\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}}{\sqrt{\hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right] \boldsymbol{\Sigma} \mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right] \hat{\boldsymbol{\mu}}}} \right) \\ & + \pi_1 \Phi \left( - \frac{\hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right] (\boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}}{\sqrt{\hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right] \boldsymbol{\Sigma} \mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right] \hat{\boldsymbol{\mu}}}} \right) \end{aligned} \quad (3.15)$$

The effect of the ensemble term,  $\mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right]$ , in (3.15) with respect to basic plug-in LDA (whose exact error has the same form as (3.15) with  $\mathbb{E}_{\mathbf{R}} \left[ \hat{\Sigma}_{\mathbf{R}}^{-1} \right]$  replaced with  $\hat{\Sigma}^{-1}$ ) is obscured by the fact that it involves the random projection  $\mathbf{R}$ . To be able to observe the effect of this term on the probability of misclassification, we derive the DE,  $\bar{\varepsilon}_{M=\infty}(\mathcal{T})$ , of the infinite discriminant-averaging RP-LDA ensemble classifier (3.7) *with respect to the random projection only*. This means that the

DE deterministically expresses the effect of the random projection ensemble for a given training set,  $\mathcal{T}$ . The result is presented in the following theorem adapted from (Niyazi et al., 2020b).

**Theorem 3.6.1.** (Infinite discriminant-averaging RP-LDA ensemble probability of misclassification DE with respect to random projection) Under the growth regime defined by the conditions (a)-(g), the following asymptotic convergence holds

$$\varepsilon_{M=\infty}(\mathcal{T}) - \bar{\varepsilon}_{M=\infty}(\mathcal{T}) \xrightarrow{\text{a.s.}} 0,$$

where

$$\begin{aligned} \bar{\varepsilon}_{M=\infty}(\mathcal{T}) := & \pi_0 \Phi \left( \frac{\hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}}{\sqrt{\hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \boldsymbol{\Sigma} \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}}} \right) \\ & + \pi_1 \Phi \left( - \frac{\hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \left( \boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}}{\sqrt{\hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \boldsymbol{\Sigma} \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}}} \right), \end{aligned} \quad (3.16)$$

where the quantity  $\hat{\nu}$  is as defined by (A.34) in Appendix A.5.

*Proof.* See Appendix A-B of (Niyazi et al., 2020b).  $\square$

Now compare the asymptotic probability of misclassification of Theorem 3.6.1 to the exact testing error of an R-LDA classifier of the form (2.1) with discriminant

$$W_{\text{R-LDA}}(\mathbf{x}, \gamma) := \hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p \right)^{-1} \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0},$$

and a threshold of  $\zeta = 0$ . The exact testing error of this classifier for a given

training set,  $\mathcal{T}$ , is

$$\begin{aligned} \varepsilon_{\text{R-LDA}}(\mathcal{T}) := & \pi_0 \Phi \left( \frac{\hat{\boldsymbol{\mu}}^T (\hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)^{-1} (\boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}}{\sqrt{\hat{\boldsymbol{\mu}}^T (\hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)^{-1} \boldsymbol{\Sigma} (\hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)^{-1} \hat{\boldsymbol{\mu}}}} \right) \\ & + \pi_1 \Phi \left( - \frac{\hat{\boldsymbol{\mu}}^T (\hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)^{-1} (\boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}}{\sqrt{\hat{\boldsymbol{\mu}}^T (\hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)^{-1} \boldsymbol{\Sigma} (\hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p)^{-1} \hat{\boldsymbol{\mu}}}} \right) \end{aligned} \quad (3.17)$$

By comparing (3.16) and (3.17) it follows that the infinite discriminant-averaging RP-LDA ensemble classifier asymptotically behaves as an R-LDA classifier with regularization parameter set to  $\gamma = \frac{1}{\hat{\nu}}$ . Since  $\hat{\nu}$  is a function of the projection dimension,  $d$ , through (A.34), the effect of  $d$  is to control this regularization parameter. Since  $d$  is an *integer* parameter, the possible values to which  $d$  can be set restrict the possible values of  $\frac{1}{\hat{\nu}}$  to a subset of  $(0, \infty)$ , whereas an R-LDA classifier's regularization parameter can vary over all of  $(0, \infty)$ . From this, we deduce that though the ensemble may be more computationally efficient than the R-LDA classifier (due to working with data of reduced dimension), for data distributed as (2.3), it asymptotically can never surpass the accuracy of an R-LDA classifier for which the regularization parameter has been properly tuned.

Not only is *invcov* a special case of the R-LDA classifier when embedded in the LDA classifier in the infinite discriminant-averaging RP-LDA ensemble classifier as we have just shown, but as is discussed in the next chapter, this estimator is in fact an instance of the more general class of *rotationally-invariant estimators*. This brings us to the subject of the next chapter: rotationally-invariant estimators of the precision matrix within the context of LDA classification.

## Chapter 4

### General Shrinkage for LDA Classification

As shown in the previous chapter, the Marzetta estimator of the precision matrix,  $invcov$ , as it appears in the RP-LDA discriminant-averaging infinite ensemble, asymptotically behaves as a regularization of the sample covariance matrix. Besides its role in classification,  $invcov$  as a standalone estimator has been shown to have the same eigenvectors as the sample covariance computed from the same data (Marzetta et al., 2011). This puts it into the category of rotationally-invariant estimators which improve upon the sample covariance estimator through a modification of the sample eigenvalues commonly referred to as ‘shrinkage’. The fact that  $invcov$ ’s shrinkage was not designed with classification in mind begs the question: why not design the optimal rotationally-invariant estimator from scratch rather than resorting to projection? Motivated by this, this chapter goes on to propose a more general non-linear form of shrinkage specifically for LDA classification.

#### 4.1 Background

Before we begin, we redefine the sample statistics introduced in Chapter 2 using an alternative notation. The pooled sample covariance estimator of  $\Sigma$  based on the two-class, common covariance data in the matrices  $\mathbf{X}_0$  and  $\mathbf{X}_1$  is denoted as

$$\hat{\Sigma}(\mathbf{X}_0, \mathbf{X}_1) := \frac{(n_0 - 1)\hat{\Sigma}(\mathbf{X}_0) + (n_1 - 1)\hat{\Sigma}(\mathbf{X}_1)}{n_0 + n_1 - 2}, \quad (4.1)$$

where  $\hat{\Sigma}(\mathbf{X}_i) := \frac{1}{n_i - 1} (\mathbf{X}_i - \hat{\boldsymbol{\mu}}(\mathbf{X}_i) \mathbf{1}_{n_i}^T) (\mathbf{X}_i - \hat{\boldsymbol{\mu}}(\mathbf{X}_i) \mathbf{1}_{n_i}^T)^T$ ,  $\hat{\boldsymbol{\mu}}(\mathbf{X}_i) := \frac{1}{n_i} \mathbf{X}_i \mathbf{1}_{n_i}$ , and  $\hat{\pi}_i := \frac{n_i}{n}$ ,  $i = 0, 1$ , are the individual sample covariance, sample mean, and

prior probability estimators, for each of the classes  $\mathcal{C}_0$  and  $\mathcal{C}_1$ , respectively. Let the eigendecomposition of the pooled sample covariance estimator be  $\hat{\Sigma}(\mathbf{X}_0, \mathbf{X}_1) = \mathbf{U}\mathbf{L}\mathbf{U}^T$ , where  $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_p]$  is the orthogonal matrix of sample eigenvectors,  $\{\mathbf{u}_j\}_{j=1}^p$ , and  $\mathbf{L} = \text{diag}(\mathbf{l})$  is the diagonal matrix of sample eigenvalues,  $\mathbf{l} = [l_1, \dots, l_p]^T$ , ordered such that  $0 \leq l_1 \leq l_2 \leq \dots \leq l_p$ . It is known that the sample covariance estimator (4.1) tends to produce biased estimates of the population eigenvalues which lead to a spreading of the sample spectrum as compared to the population spectrum (Friedman, 1989). These effects worsen with decreasing sample size. To counteract this phenomenon, Charles Stein proposed the first instance of an estimator belonging to the class of *rotationally-invariant estimators* which directly modify the distorted sample eigenvalues (Stein, 1986).

A rotationally-invariant estimator of the covariance/precision matrix, denoted by  $\tilde{\Sigma}(\mathbf{X}_0, \mathbf{X}_1)$ , is an estimator for which, when the underlying data is rotated, the estimator is rotated in the same way (Ledoit and Wolf, 2012). Mathematically, this is expressed as

$$\tilde{\Sigma}(\mathbf{W}\mathbf{X}_0, \mathbf{W}\mathbf{X}_1) = \mathbf{W}\tilde{\Sigma}(\mathbf{X}_0, \mathbf{X}_1)\mathbf{W}^T, \quad (4.2)$$

where  $\mathbf{W} \in \mathbb{R}^{p \times p}$  is an orthogonal matrix. It is easy to see that the sample covariance estimator  $\hat{\Sigma}(\mathbf{X}_0, \mathbf{X}_1)$  itself satisfies (4.2) and thus belongs to the class of rotationally-invariant estimators. In fact, since the eigenvectors of  $\hat{\Sigma}(\mathbf{W}\mathbf{X}_0, \mathbf{W}\mathbf{X}_1)$  are  $\mathbf{W}\mathbf{U}$ , any estimator which keeps the sample covariance eigenvectors as its eigenvectors and modifies only the sample eigenvalues is rotationally-invariant. Such estimators can be expressed in the (unrotated) form

$$\tilde{\Sigma}(\mathbf{X}_0, \mathbf{X}_1) = \mathbf{U}\mathbf{D}\mathbf{U}^T, \quad (4.3)$$

with  $\mathbf{D} = \text{diag}(d_1, \dots, d_p)$ , where  $d_j := \phi(l_j)$ ,  $j = 1, \dots, p$ , and  $\phi: \mathbb{R}_+ \cup \{0\} \rightarrow \mathbb{R} - \{0\}$  is called the *shrinkage function*. Over the years, a multitude of shrinkages have been proposed for a variety of different metrics. The classical Stein estimator

designs the shrinkage to minimize a metric which is equivalent to the Kullback-Leibler divergence (or relative entropy) of the distribution  $\mathcal{N}(\boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}(\mathbf{X}_0, \mathbf{X}_1))$  from  $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , up to a factor of 1/2. In a succession of more than a dozen papers, Ledoit and Wolf propose linear and non-linear shrinkage estimators of the sample covariance matrix (Ledoit and Wolf, 2022a). The common theme of these works is the optimization of the shrinkages based on the Frobenius norm of the difference between the true covariance and the estimator. Needless to say, for an application such as the sample covariance estimation for classification considered in this paper, the Stein and Ledoit and Wolf estimator loss functions result in non-optimal performance as compared to estimators designed for more relevant metrics like the probability of misclassification, true positive rate, false positive rate, etc...

An example of a commonly-used shrinkage function in the context of LDA classification, is the regularization of the sample covariance matrix of the form

$$\hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p, \quad (4.4)$$

where  $\gamma > 0$  is a parameter which is tuned according to the metric of interest (Guo et al., 2007). It is easy to see that the regularized sample covariance matrix,  $\hat{\boldsymbol{\Sigma}}(\mathbf{X}_0, \mathbf{X}_1) + \gamma \mathbf{I}_p = \mathbf{U}(\mathbf{L} + \gamma \mathbf{I}_p) \mathbf{U}^T$ , has the same eigenvectors as the sample covariance matrix, but shifted eigenvalues  $l_1 + \gamma, l_2 + \gamma, \dots, l_p + \gamma$ . For LDA, these are then inverted to form an estimate of the precision matrix so that the final shrinkages are of the form  $\frac{1}{l_1 + \gamma}, \frac{1}{l_2 + \gamma}, \dots, \frac{1}{l_p + \gamma}$ . As we have seen, another rotationally-invariant estimator which has been used in the context of LDA classification (yielding the discriminant-averaging RP-LDA ensemble classifier of (Durrant and Kabán, 2015)) is the Marzetta estimator, ‘*invcov*’, of the precision matrix based on random projections (Marzetta et al., 2011). Marzetta et al. (2011) shows that *invcov*’s particular shrinkage shifts all zero eigenvalues of the sample covariance to a non-zero constant while modifying the non-zero eigenvalues non-trivially, however, as presented in Chapter 3, asymptotic analy-

sis shows that the shrinkage is effectively of the form  $d_j = \frac{1}{l_j + \gamma(d)}$ ,  $j = 1, \dots, p$ , where  $\gamma(d) > 0$  is a constant which depends on the chosen dimension,  $d$ , of the random projections. This result is also proven in Appendix B.2.1 for *invcov* in isolation. This means that *invcov* is essentially an inverted regularized sample covariance like (4.4), but with coarser tuning of its regularization parameter since  $\gamma(d)$  varies only through the *integer* parameter  $d$ . So, while these two estimators allow for customization to relevant metrics in classification through their respective hyperparameters, they are restricted in the sense that, prior to inversion, the shrinkage is linear in the sample eigenvalues.

Finally, references (Sifaou et al., 2020) and (Li et al., 2022) propose non-linear shrinkage in order to minimize LDA misclassification rate under a spiked covariance model. More specifically, these papers assume a covariance of a scaled identity matrix plus a finite rank perturbation. This model gives rise to a sample spectrum composed of a finite number of zero probability mass spikes surrounding a bulk. Reference (Sifaou et al., 2020) assumes that the spikes are only to the right of the bulk, while reference (Li et al., 2022) allows the spikes to appear on both sides of the bulk. Since the optimal shrinkages corresponding to the bulk are known in advance, the spiked model assumption reduces the number of shrinkages to be determined to a finite number corresponding to the number of spikes. While this model does have a number of specialized applications, it is generally restrictive. This leads us to our proposed shrinkage which is both non-linear, in general, and makes no assumptions on the structure of the covariance, while utilizing a finite number of parameters.

## 4.2 Contributions

With the above goal in mind, and with the constraint of requiring a shrinkage that can be consistently estimated in order to be useful in practice, we propose a form of shrinkage composed of a weighted linear combination of sub-component shrinkage function. Placing the resulting rotationally-invariant estimator in place of the

precision matrix estimator in the LDA decision rule gives rise to the Weighted-Shrinkage Linear Discriminant Analysis (WS-LDA) classifier. We provide a G-estimator of the probability of misclassification of WS-LDA as well as that of any shrinkage with LDA base classifier under certain conditions. Additionally, we optimize the classifier's hyperparameters. Simulations are performed on synthetic data in order to compare the proposed classifier with other rotationally-invariant schemes such as R-LDA and several of Ledoit and Wolf's estimators as different parameters of the classification problem are varied. The contributions of this chapter are summarized as

- A general non-linear rotationally-invariant estimator designed specifically for LDA classification whose shrinkage is a weighted linear combination of a fixed set of sub-component shrinkage functions. We call the resulting classifier the WS-LDA classifier.
- A G-estimator of the probability of misclassification of WS-LDA for any setting of hyperparameters which, as a special case, applies to any shrinkage under some conditions on the shrinkage functions
- The optimal setting of hyperparameters of WS-LDA which minimize the G-estimator of the probability of misclassification.

In the remainder of this chapter, we drop the arguments of the sample statistic estimators defined at the beginning of Section 4.1. Hereinafter, the pooled sample covariance,  $\hat{\Sigma}(\mathbf{X}_0, \mathbf{X}_1)$ , is denoted by  $\hat{\Sigma}$ , the rotationally-invariant estimator based on the same data,  $\tilde{\Sigma}(\mathbf{X}_0, \mathbf{X}_1)$ , is denoted by  $\tilde{\Sigma}$ , the sample covariance estimate for data in  $\mathcal{C}_i$ ,  $\hat{\Sigma}(\mathbf{X}_i)$ , is denoted by  $\hat{\Sigma}_i$ , and, finally, the sample mean corresponding to class  $\mathcal{C}_i$ ,  $\hat{\mu}(\mathbf{X}_i)$ , is denoted by  $\hat{\mu}_i$ ,  $i = 0, 1$ .

### 4.3 Rotationally-invariant estimators in the context of LDA classification

#### 4.3.1 The rotationally-invariant LDA Rule

Consider the LDA discriminant

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}, \quad (4.5)$$

which classifies the test point  $\mathbf{x}$  to class  $\mathcal{C}_1$  when (4.5) is positive, and to class  $\mathcal{C}_0$  otherwise. This rule simply plugs in the maximum likelihood estimates of the class statistics into the Bayes rule for Gaussian data with common covariance.

Now consider the alternative discriminant

$$\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}} \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0}, \quad (4.6)$$

where  $\tilde{\boldsymbol{\Sigma}}$  is understood to be a rotationally-invariant estimator of the **precision** matrix. As a special case, setting  $\tilde{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}^{-1}$  recovers (4.5). Moreover,  $\tilde{\boldsymbol{\Sigma}}$  may be set to any of the aforementioned estimators, with inversion where appropriate. Note, however, that depending on the chosen  $\tilde{\boldsymbol{\Sigma}}$ , this approach may end up designing the estimator independently from the classifier. For example, if the Stein or Ledoit and Wolf estimators of the sample covariance are employed, they are simply inverted and plugged in.

To further emphasize this point, we now derive the (unobservable) Bayes shrinkage and show that it is not only a function of the true covariance matrix, but also the class means. This implies that for optimal performance it is not enough to design a perfect precision matrix estimator in isolation; the context of the estimator must also be taken into account.

### 4.3.2 Bayes shrinkage

The Bayes shrinkage for our data assumptions is the shrinkage which results in the minimum error for Gaussian classes with common covariance, or, equivalently, it is the shrinkage yielding a weight vector which is proportional to the Bayes weight vector. As is to be expected, the Bayes shrinkage depends on a knowledge of both the true means and true covariance, as we will show in what follows.

First, the Bayes' weight vector is  $\mathbf{w}_{\text{Bayes}} := \Sigma^{-1}\boldsymbol{\mu}$ . In fact, any vector which is proportional to  $\mathbf{w}_{\text{Bayes}}$  will produce the same misclassification error. Based on this, the design weight vector for known means based on the rotationally-invariant estimator (4.3) is then  $\mathbf{UDU}\boldsymbol{\mu}$ . This can be expressed as  $\mathbf{UDU}\boldsymbol{\mu} = \sum_{j=1}^p d_j \mathbf{u}_j \mathbf{u}_j^T \boldsymbol{\mu}$ . To determine the Bayes shrinkage,  $d_j^{\text{Bayes}}$ ,  $j = 1, \dots, p$ , we need to express the Bayes weight vector in the form of this design expression.

To do this, project  $\mathbf{w}_{\text{Bayes}}$  onto the sample eigenvectors (which form a basis). Then,  $\mathbf{w}_{\text{Bayes}} = \sum_{j=1}^p \beta_j \mathbf{u}_j$ , where  $\beta_j = \frac{\mathbf{u}_j^T \Sigma^{-1} \boldsymbol{\mu}}{\mathbf{u}_j^T \mathbf{u}_j} = \mathbf{u}_j^T \Sigma^{-1} \boldsymbol{\mu}$ . So we have,

$$\begin{aligned} \mathbf{w}_{\text{Bayes}} &= \sum_{j=1}^p \beta_j \mathbf{u}_j \\ &= \sum_{j=1}^p \frac{\beta_j}{\mathbf{u}_j^T \boldsymbol{\mu}} \mathbf{u}_j \mathbf{u}_j^T \boldsymbol{\mu} \\ &= \sum_{j=1}^p \frac{\mathbf{u}_j^T \Sigma^{-1} \boldsymbol{\mu}}{\mathbf{u}_j^T \boldsymbol{\mu}} \mathbf{u}_j \mathbf{u}_j^T \boldsymbol{\mu}, \end{aligned}$$

from which we observe that the shrinkage corresponding to  $\mathbf{w}_{\text{Bayes}}$  is  $d_j^{\text{Bayes}} := \frac{\mathbf{u}_j^T \Sigma^{-1} \boldsymbol{\mu}}{\mathbf{u}_j^T \boldsymbol{\mu}}$ ,  $j = 1, \dots, p$ . Note that this shrinkage is of the form  $\phi(\mathbf{u}_j, \Sigma, \boldsymbol{\mu})$ ,  $j = 1, \dots, p$ , where we expected a shrinkage of the form  $\phi(l_j)$ ,  $j = 1, \dots, p$ . For a proof that this shrinkage satisfies the criteria for rotational-invariance in (4.2), see Appendix B.2.2. This result is important in that it shows that the Bayes weight vector is theoretically achievable within the class of rotationally-invariant estimators. Alas, it is difficult to actually estimate  $d_j^{\text{Bayes}}$  without making any further assumptions on  $\Sigma$ . In any case, this result motivates the use of rotationally-

invariant estimators in the context of LDA classification as done in this work.

Although unattainable in practice, the Bayes shrinkage can be used as a benchmark against which to compare other shrinkages in the synthetic data simulations in Section 4.5. More importantly, it shows that in order to even approach the optimal shrinkage for LDA, (4.3) must depend on some estimate of the class means in addition to the covariance. In the next section, we move in this direction by proposing a form of shrinkage for which consistent estimation is tractable.

## 4.4 Main results

This section restricts the form of shrinkage considered in this work to a weighted linear combination of sub-component shrinkage functions ensuring the possibility of consistent estimation of the shrinkage parameters. From there, it derives the corresponding G-estimator of the probability of misclassification of the LDA classifier based on this shrinkage and the optimal weights and classifier intercept for those weights.

### 4.4.1 Proposed form of shrinkage

In this work, we consider a rotationally-invariant estimator whose shrinkage takes the form of a weighted sum of a finite number of sub-component shrinkage functions. As such, we call it the Weighted Shrinkage (WS) estimator defined as follows

$$\tilde{\Sigma}_{\text{WS}} = \mathbf{U}\mathbf{D}_{\text{WS}}\mathbf{U}^T, \quad (4.7)$$

with  $\mathbf{D}_{\text{WS}} = \text{diag}(d_1^{\text{WS}}, \dots, d_p^{\text{WS}})$ , where

$$\begin{aligned} d_j^{\text{WS}} &:= \sum_{k=1}^r \alpha_k h_k(l_j) \\ &= \boldsymbol{\alpha}^T \mathbf{h}(l_j), \quad j = 1, \dots, p, \end{aligned}$$

where  $\{h_k\}_{k=1}^r$  are a set of  $r$  functions of the sample eigenvalues and  $\{\alpha_k\}_{k=1}^r$  are their coefficients. The vectors  $\mathbf{h}(l_j)$  and  $\boldsymbol{\alpha}$  are then defined as  $\mathbf{h}(x) = [h_1(x), \dots, h_r(x)]^T$  and  $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_r]^T$ . Note that this form is able to accommodate any shrinkage as a special case when  $r = 1$ ,  $\alpha_1 = 1$ , and  $h_1$  is set to that particular shrinkage function. Also note that this form reduces the number of parameters in (4.3) from  $p$  (shrinkages) to  $r$  (coefficients). This is important because, in general, it is impossible to consistently estimate  $p$  parameters from  $n$  observations when  $n$  is on the order of  $p$ . The number of coefficients,  $r$ , however, is assumed to be finite with respect to  $p$  and  $n$ , thereby making the estimation tractable.

With the rotationally-invariant estimator (4.7) in hand, we now consider the classifier with the discriminant

$$\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta. \quad (4.8)$$

Here,  $\boldsymbol{\alpha}$  and  $\theta$  are hyperparameters of the classifier. The exact probability of misclassification corresponding to this classifier is presented in the following lemma.

**Lemma 4.4.1.** The exact probability of misclassification of the classifier (4.8) under the data assumption (2.3) for a given training set is

$$\pi_0 \Phi \left( \frac{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}}}} \right) + \pi_1 \Phi \left( -\frac{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}}}} \right). \quad (4.9)$$

*Proof:* The proof is similar to that of Lemma 1 in (Niyazi et al., 2020a).

The main contribution of this work is to provide the optimal coefficients,  $\boldsymbol{\alpha}$ , and intercept,  $\theta$ , which minimize (4.9) for any given set of sub-component shrinkage functions,  $\{h_k\}_{k=1}^r$ . To do this, we first derive an estimator of (4.9) which is consistent when  $p$  and  $n$  grow at constant rates to each other for a fixed number of sub-component shrinkage functions,  $r$ . We then optimize this

expression jointly over  $\boldsymbol{\alpha}$  and  $\theta$ . This is detailed in the next section. By standard arguments related to the uniformity of a function over its parameters (see, for example, (Yang et al., 2018)), this is asymptotically equivalent to optimizing over the exact probability of misclassification, (4.9).

#### 4.4.2 Assumptions and main results

In this section, we derive the G-estimator of (4.9) under standard RMT assumptions as well as other assumptions specific to the LDA classification context and the form of shrinkage imposed in (4.7). Theorem 4.4.1 presents this G-estimator. Building on this, Theorem 4.4.2 then provides the optimal  $\boldsymbol{\alpha}$  and  $\theta$  which jointly minimize the G-estimator.

The assumptions for the validity of Theorems 1 and 2 are:

- (a)  $p$  and  $n$  grow together such that  $0 < \liminf \frac{p}{n} < \limsup \frac{p}{n} < \infty$ ;
- (b)  $\frac{r}{p} \rightarrow 0$  and  $\frac{r}{n} \rightarrow 0$ , i.e.,  $r$  is fixed in relation to  $p$  and  $n$ ;
- (c)  $0 < \liminf_p \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 < \limsup_p \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 < \infty$ ;
- (d)  $\limsup_p \|\boldsymbol{\Sigma}\|_2 < \infty$ ;
- (e)  $\liminf_p \lambda_{\min}(\boldsymbol{\Sigma}) > 0$ ;
- (f) the set of shrinkages,  $\{h_k\}_{k=1}^r$ , are analytical and bounded in an open set of  $\mathbb{C}$  containing the interval  $[-\epsilon, \infty)$ ,  $\epsilon > 0$ ;
- (g)  $\|\boldsymbol{\alpha}\|_2 > 0$ ;

Now, let  $\rho := \text{rank} \left\{ \hat{\boldsymbol{\Sigma}} \right\}$  and define

$$\bar{\mathbf{U}} := [\mathbf{u}_{p-\rho+1} \cdots \mathbf{u}_p], \quad (4.10)$$

$$\bar{\mathbf{I}} := [l_{p-\rho+1} \cdots l_p]^T, \quad (4.11)$$

$$\bar{\mathbf{L}} := \text{diag}(\bar{\mathbf{I}}), \quad (4.12)$$

$$\bar{\boldsymbol{\Sigma}} := \bar{\mathbf{U}} \bar{\mathbf{L}} \bar{\mathbf{U}}^T, \quad (4.13)$$

$$\bar{\mathbf{Q}}(z) := (\bar{\boldsymbol{\Sigma}} - z \mathbf{I}_\rho)^{-1}, \quad (4.14)$$

and  $\{\nu_i\}_{i=1}^\rho$  as the eigenvalues of  $\bar{\mathbf{L}} - \left(\sqrt{\frac{\bar{\mathbf{I}}}{n-2}}\right) \left(\sqrt{\frac{\bar{\mathbf{I}}}{n-2}}\right)^T$ . Note that if  $\rho = p$ , then (4.10) is simply  $\mathbf{U}$ , (4.11) is simply  $\mathbf{l}$ , (4.12) is simply  $\mathbf{L}$ , and (4.13) is simply  $\hat{\Sigma}$ .

**Theorem 4.4.1.** The G-estimator of the probability of misclassification (4.9) of the classifier (4.8) is

$$\sum_{i=0}^1 \hat{\pi}_i \Phi \left( (-1)^i \frac{\boldsymbol{\alpha}^T \left[ \frac{(-1)^{i+1}}{2} \mathbf{H}_1 \mathbf{b} + (-1)^i \mathbf{H}_\nu \mathbf{c}_i \right] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \right), \quad (4.15)$$

where  $\mathbf{b} := \mathbf{U}^T \hat{\boldsymbol{\mu}} \circ \mathbf{U}^T \hat{\boldsymbol{\mu}}$ ,  $\mathbf{c}_i := \left[ \frac{\frac{1}{n_i-1} \text{tr}\{\bar{\Sigma} \bar{\mathbf{Q}}(\nu_1)\}}{\frac{1}{n-2} \text{tr}\{\bar{\Sigma} \bar{\mathbf{Q}}^2(\nu_1)\}}, \dots, \frac{\frac{1}{n_i-1} \text{tr}\{\bar{\Sigma} \bar{\mathbf{Q}}(\nu_\rho)\}}{\frac{1}{n-2} \text{tr}\{\bar{\Sigma} \bar{\mathbf{Q}}^2(\nu_\rho)\}} \right]^T$ ,  $[\mathbf{\Pi}]_{l,k} := \frac{\hat{\boldsymbol{\mu}}^T \bar{\mathbf{Q}}(\nu_l) \bar{\Sigma} \bar{\mathbf{Q}}(\nu_k) \hat{\boldsymbol{\mu}}}{\frac{1}{n-2} \text{tr}\{\bar{\Sigma} \bar{\mathbf{Q}}^2(\nu_l)\} \frac{1}{n-2} \text{tr}\{\bar{\Sigma} \bar{\mathbf{Q}}^2(\nu_k)\}}$ ,  $l, k = 1, \dots, \rho$ , the  $r \times p$  matrix  $\mathbf{H}_1 := [\mathbf{h}(l_1) \cdots \mathbf{h}(l_p)]$ , the  $r \times \rho$  matrix  $\mathbf{H}_\nu := [\mathbf{h}(\nu_1) \cdots \mathbf{h}(\nu_\rho)]$ , and it is assumed that  $\liminf_p \lambda_{\min}(\mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T) > 0$  is satisfied.

*Proof:* See Appendix B.1.1.

This result is useful as it provides us with the G-estimator of the probability of misclassification of (4.8) for any combination of shrinkage and intercept as long as the shrinkage is of the form (4.7) whose sub-component shrinkage functions satisfy assumption (f) and for which  $r$  is small compared to the data dimensions  $p$  and  $n$ .

The next theorem presents the optimal coefficients,  $\boldsymbol{\alpha}^*$ , and optimal intercept,  $\theta^*$ , which jointly minimize (4.15) for a classifier of the form (4.8). The corresponding G-estimator for this optimal combination of coefficients and intercept is also presented.

**Theorem 4.4.2.** Assuming that  $\mathcal{C}_0$  is the majority class, the optimal coefficients,  $\boldsymbol{\alpha}^*$ , and intercept,  $\theta^*$ , which jointly minimize (4.15) along with the optimal intercept, are

$$\boldsymbol{\alpha}^* = (\mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T)^{-1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}], \quad (4.16)$$

where  $\mathbf{c} = \mathbf{c}_0 + \mathbf{c}_1$ , and

$$\theta^* = -\ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \frac{\boldsymbol{\alpha}^{*T} \mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}^*}{\boldsymbol{\alpha}^{*T} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \frac{\boldsymbol{\alpha}^{*T} \mathbf{H}_\nu (\mathbf{c}_1 - \mathbf{c}_0)}{2},$$

with corresponding G-estimator of the probability of misclassification

$$\hat{\pi}_0 \Phi \left( -\frac{1}{2} \tau^* + \frac{1}{\tau^*} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) + \hat{\pi}_1 \Phi \left( -\frac{1}{2} \tau^* - \frac{1}{\tau^*} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right), \quad (4.17)$$

where  $\tau^* := \frac{\boldsymbol{\alpha}^{*T} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]}{\sqrt{\boldsymbol{\alpha}^{*T} \mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}^*}}$ .

*Proof:* See Appendix B.1.2.

Theorem 4.4.2 is useful as it allows us to construct a classifier of the form (4.8) which minimizes the G-estimator of the probability of misclassification over the hyperparameters  $\boldsymbol{\alpha}$  and  $\theta$ . This is achieved by plugging  $\boldsymbol{\alpha}^*$  and  $\theta^*$  into (4.8). Theorem 4.4.2 also provides us with the G-estimator of the probability of misclassification of this optimal classifier (4.17). Additionally, (4.17) is able to recover the G-estimator of the probability of misclassification of (4.8) with any linear combination of shrinkage functions and corresponding optimal intercept by setting  $\boldsymbol{\alpha}^*$  to an  $\boldsymbol{\alpha}$  of choice. In particular, setting  $r = 1$ ,  $\alpha_1 = 1$ , and  $h_1$  to a particular shrinkage which satisfies (f) yields the G-estimator of the probability misclassification of (4.8) for any given shrinkage function (R-LDA, for example) and optimal intercept. In any case, when  $r = 1$ , (4.16) is in fact a scalar whose effect cancels out when computing  $\tau^*$  and so the coefficient,  $\alpha_1$ , no longer matters.

## 4.5 Simulations

This section compares the performance of the proposed classifier (4.8), with a given set of sub-component shrinkage functions and their optimal coefficients, to several popular rotationally-invariant estimators in the literature which are combined with LDA in the form (4.6). The simulation design closely mirrors that of (Ledoit and Wolf, 2022b) with a baseline scenario of synthetic data for which

certain parameters are varied and the performance of each classifier is plotted in each case. The baseline scenario we consider in this paper is as follows:

- the data dimension is  $p = 200$ ;
- the training set sample size is  $n = 600$  so that the baseline concentration ratio is  $p/n = 1/3$ ;
- the training data is Gaussian following the distribution (2.5);
- the class means are  $\boldsymbol{\mu}_0 = \frac{1}{p^{1/4}} \left[ 3 \times \mathbf{1}_{\lceil \sqrt{p} \rceil}^T \mathbf{0}_{p - \lceil \sqrt{p} \rceil - 2}^T 5 \ 5 \right]^T$  and  $\boldsymbol{\mu}_1 = \mathbf{0}_p$ ;
- the default condition number of the population covariance matrix is 30;
- the population covariance matrix is diagonal with 20% of the eigenvalues equal to 1, 40% of the eigenvalues equal to 3, and 40% of the eigenvalues equal to 10;
- and the class priors are equal.

The rotationally-invariant estimators considered in the context of the LDA classifier are:

- the precision matrix estimator,  $\hat{\boldsymbol{\Sigma}}^{-1}$ , which yields the LDA classifier (4.5);
- the regularized and subsequently inverted sample covariance matrix estimator,  $\left( \hat{\boldsymbol{\Sigma}} + \gamma \mathbf{I}_p \right)^{-1}$ , which yields R-LDA, where  $\gamma$  is a hyperparameter to be tuned;
- Ledoit and Wolf's analytical estimator which is non-linear and optimizes the Frobenius norm (Ledoit and Wolf, 2020);
- and Ledoit and Wolf's Quadratic Inverse Shrinkage (QIS) estimator which is also non-linear and also optimizes the Frobenius norm (Ledoit and Wolf, 2022b).

The above schemes are compared to the proposed WS-LDA classifier (4.8) with  $r = 6$ ,  $h_1(x) = 1$ ,  $h_2(x) = \frac{1}{x+\gamma}$ ,  $h_3(x) = \frac{1}{(x+\gamma)^2}$ ,  $h_4(x) = \frac{1}{(x+\gamma)^3}$ ,  $h_5(x) = \frac{1}{(x+\gamma)^4}$ , and  $h_6(x) = \frac{1}{(x+\gamma)^5}$ , where  $\gamma$  is the same as the regularization parameter of R-LDA and  $\alpha$  is set to the optimal  $\alpha^*$  in (4.16). For fair comparison, all classifier intercepts are set to their optimal intercept (B.25). The hyperparameter  $\gamma$  is tuned according to the G-estimator (4.17) of the R-LDA probability of misclassification. The exact testing error for a given training set for each classifier is evaluated using (4.9). This quantity is then averaged over 1000 training sets in order to estimate the expected testing error. We then report the Percentage Relative Improvement in Error (PRIE) defined similarly to the Percentage Relative Improvement in Average Loss (PRIAL) in (Ledoit and Wolf, 2022b), but based on the expected testing error. The PRIE is defined for a rotationally-invariant estimator,  $\tilde{\Sigma}$ , as

$$\text{PRIE}(\tilde{\Sigma}) = \frac{\varepsilon(\hat{\Sigma}^{-1}) - \varepsilon(\tilde{\Sigma})}{\varepsilon(\hat{\Sigma}^{-1}) - \varepsilon_{\text{Bayes}}} \times 100,$$

where  $\varepsilon(\hat{\Sigma}^{-1})$  is the expected error corresponding to LDA as in (4.5),  $\varepsilon(\tilde{\Sigma})$  is the expected testing error corresponding to LDA with a particular shrinkage as in (4.6), and  $\varepsilon_{\text{Bayes}}$  is the Bayes error. The PRIE signifies the improvement of  $\tilde{\Sigma}$  in the context of LDA relative to the sample covariance matrix as a percentage of the improvement of the Bayes shrinkage relative to the sample covariance matrix. By definition, its magnitude is no more than 1. Positive PRIE values indicate that  $\tilde{\Sigma}$  improves performance relative to the sample covariance matrix, while negative PRIE values indicate that performance is degraded by  $\tilde{\Sigma}$  relative to the sample covariance matrix. An important point to note here is that we use exact errors which we have access to since we use synthetic data, but even if we were to use a training data based error estimator such as cross-validation to evaluate the performance of each classifier, tuning any classifier hyperparameters using the G-estimator based on the same training data does not introduce any selection bias, as the G-estimator is a consistent estimator of the true probability

of misclassification. There is therefore no need to resort to computationally-draining procedures such as nested cross-validation in order to avoid selection bias due to overlap between the data used for training and tuning and the data used for performance evaluation.

Lastly, a final remark concerning the inversion of Ledoit and Wolf’s analytical and QIS sample covariance estimators to obtain estimators for the precision matrix is that, although doing this is less optimal in terms of the Frobenius norm than if a precision matrix estimator had been designed directly, it is not totally arbitrary. Indeed, both estimators share the same oracle which minimizes the inverse Stein loss (Ledoit and Wolf, 2022b). Therefore, inverting these estimators results in precision matrix estimators which asymptotically minimize the Stein’s loss for the precision matrix. Note that these estimators are implemented using the authors’ own code which is available at [https://www.econ.uzh.ch/en/people/faculty/wolf/publications.html#Programming\\_Code](https://www.econ.uzh.ch/en/people/faculty/wolf/publications.html#Programming_Code).

We now present the results of the simulations when varying the data dimensions, the condition number, the concentration ratio, and the priors.

## Convergence

In this simulation, the data dimensions  $p$  and  $n$  are made to grow at the baseline concentration ratio of  $p/n = 1/3$ . Figure 4.1 plots the PRIE against  $p$  for each of the rotationally-invariant schemes. As expected, the performance of WS-LDA improves with increasing dimension as the G-estimator of Theorem 4.4.1 based on which the classifier hyperparameters are optimized becomes more accurate. Consequently,  $\alpha^*$  becomes more accurate, resulting in a more optimal classifier and at  $p = 1000$ , WS-LDA achieves a 22.5% improvement in performance as compared to LDA. In contrast, R-LDA and the Ledoit and Wolf estimators do very poorly, with R-LDA having almost 0% improvement and the Ledoit and Wolf estimators degrading performance as compared to the sample covariance estimator.

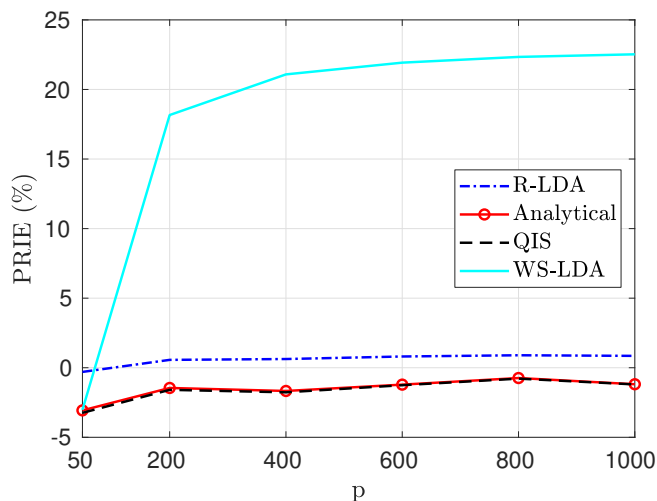


Figure 4.1: Plot of the PRIE against varying data dimension,  $p$ , where  $p$  and  $n$  grow together at a fixed ratio of  $1/3$ .

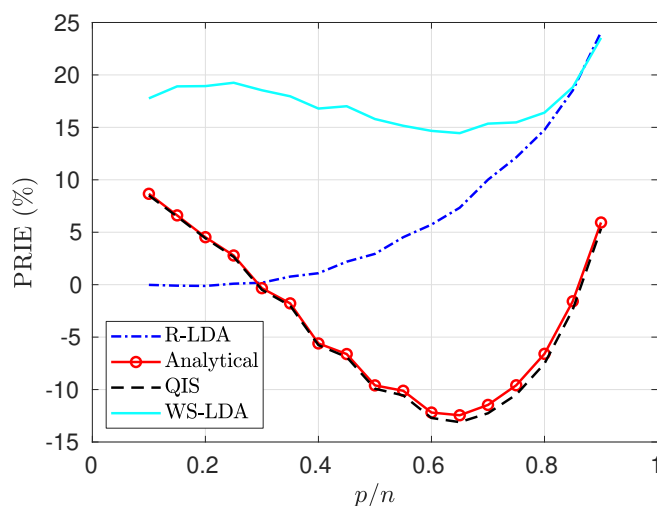


Figure 4.2: Plot of the PRIE against varying concentration ratio,  $p/n$ .

## Concentration ratio

In this simulation, the concentration ratio,  $p/n$ , is varied while keeping the product of the dimensions,  $p \times n$ , constant at the same level of the baseline of  $200 \times 600 = 120,000$ . This ensures that the amount of information in the data matrix remains constant as the concentration ratio is varied Ledoit and Wolf (2022b). Figure 4.2 plots the PRIE against the concentration ratio for each of the rotationally-invariant schemes. WS-LDA outperforms all other classifiers except at high concentration ratios at which it is matched by R-LDA. This is analogous to the behavior observed in Ledoit and Wolf (2022b); at high concen-

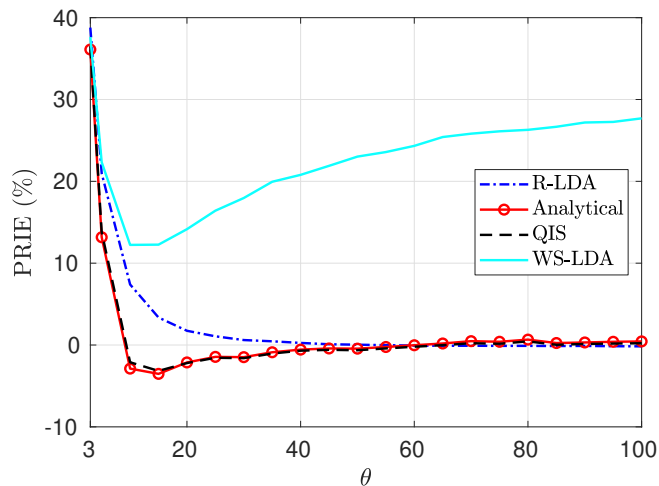


Figure 4.3: Plot of the PRIE against varying condition number,  $\theta$ .

tration ratios the PRIAL of the linear shrinkage approaches the PRIALs of the non-linear shrinkages. In terms of the PRIE, Ledoit and Wolf's estimators do relatively poorly, with degraded performance as compared to the sample covariance estimator between concentration ratios of 0.3 and 0.9.

## Condition number

This simulation investigates the effect of varying the condition number of the population covariance matrix on the PRIE of each classifier. Let  $\theta$  denote the condition number of  $\Sigma$ . Then for  $\theta > 1$  under the baseline scenario, 20% of the population eigenvalues are 1, 40% of the population eigenvalues are  $\frac{2\theta+7}{9}$ , and 40% of the population eigenvalues are  $\theta$ . Figure 4.3 plots the PRIE against the condition number for each of the rotationally-invariant schemes. At very low condition numbers in the range of  $\theta = 3$  to  $\theta = 6$  all classifiers perform similarly, however with increasing condition number, WS-LDA improves in performance while the other classifiers approach a PRIE of 0%.

## Priors

In this simulation, we vary the class priors. Figure 4.4 plots the PRIE against  $\pi_0$  for each of the rotationally-invariant schemes. Note that we vary  $\pi_0$  such that it is always the majority class, as this assumption underlies the derivation of Theorem

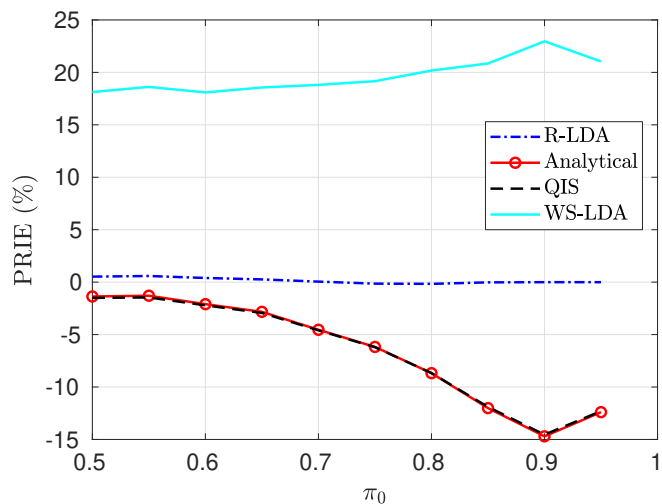


Figure 4.4: Plot of the PRIE against varying class  $\mathcal{C}_0$  prior,  $\pi_0$ .

4.4.2 (see Appendix B.1.2). WS-LDA exhibits a relatively stable PRIE of around 20% with increasing  $\pi_0$ , R-LDA stays fixed at around 0% PRIE, and the Ledoit and Wolf estimators show a gradual degradation in performance with increasing  $\pi_0$ .

This concludes the current chapter on rotationally-invariant estimator for LDA classification. The next chapter explores the broader scope of generic binary linear classifiers and how their performance may be improved through weight vector tuning.

## Chapter 5

### Weight-Vector Tuning of Linear Classifiers

Up till now, we have been looking at the analysis and design of high-dimensional variants of the classical LDA classifier. In this chapter, we broaden our scope to generic binary linear classifiers.

Unlike its intercept, a linear classifier's weight vector cannot be tuned by a simple grid search. This chapter proposes a technique for tuning the weight vector of any binary linear classifier through a scalar parameter. This is achieved by a parameterization of a decomposition of the discriminant by a scalar which controls the trade-off between conflicting informative and noisy terms. By varying this parameter, the original weight vector is modified in a meaningful way. Applying this method to a number of linear classifiers under a variety of data dimensionality and sample size settings reveals that the classification performance loss due to non-optimal native hyperparameters can be compensated for by weight vector tuning. This yields computational savings as the proposed tuning method reduces to tuning a scalar compared to tuning the native hyperparameter, which may involve repeated weight vector generation along with its burden of optimization, dimensionality reduction, etc., depending on the classifier. It is also found that weight vector tuning significantly improves the performance of LDA under high estimation noise. Proceeding from this second finding, an asymptotic study of the misclassification probability of the parameterized LDA classifier in the growth regime where the data dimensionality and sample size are comparable is conducted. Using RMT, the misclassification probability is shown to converge to a quantity that is a function of the true statistics of the data. Additionally, a G-estimator of the misclassification probability is derived and computationally

efficient tuning of the parameter using this estimator is demonstrated on real data.

The proposed weight vector tuning technique goes beyond the LDA-based variants studied in the previous chapters to allow improved performance in small sample regimes for a multitude of classifiers including the SVM and even neural networks. The latter observation paves the way for applications of the proposed method in transfer learning. This chapter is concluded with an application of the proposed technique to transfer learning in neural networks. This is possible as long as the last layer of the neural network is fully-connected, and therefore linear.

The next section provides a background on the relevant literature.

## 5.1 Background

As explained in Chapter 2, a binary linear classifier classifies a data point to one class or the other by thresholding a discriminant that is a linear combination of the data features. The weights of the features make up a *weight vector* and the constant term in the discriminant is the *bias* of the classifier.

Despite the availability of sophisticated non-linear methods for classification, linear classifiers are still widely used. In fact, new variants of standard linear methods catering to specific settings and applications are being developed all the time. A search of the recent literature reveals that linear classifiers are being employed in many tasks including clinical neuroimaging (Marquand and Kia, 2020), digital pulse shape discrimination (Wen et al., 2020), predicting the genetic merit of beef cattle (Berry et al., 2019), and in conjunction with other methods for applications such as pathogen identification (Randhawa et al., 2020), strategy representation (Ashok et al., 2019), and cancer classification (Alanni et al., 2019). Linear classifiers are especially suited to certain high-dimensional datasets on which they perform comparably with non-linear classifiers, with the advantage of much faster training times and quicker classification (Yuan et al., 2012). Due to

ease of computation, linear classifiers further make good trial classifiers during the initial exploratory phase, when the relationship between the data features and labels is yet unknown (Duda et al., 2001).

One way of improving a given linear classifier’s performance on a particular dataset is by tuning its bias so as to minimize training error on that dataset (Friedman et al., 2001). Because the bias is a scalar, a grid search for the optimum is computationally undemanding. Even the need for a grid-search can be eliminated in many cases for which explicit representations of the optimal bias can be derived. For example, Wang et al. (2018) derive an explicit bias correction of the LDA classifier discriminant in order to improve classification in the high estimation noise regime. Zollanvari et al. (2019) similarly correct for the bias of this classifier in an explicit form, but in the context of cost-sensitive classification. Additionally, the references (Huang et al., 2010) and (Sifaou et al., 2020) provide explicit bias corrections for certain high-dimensional variants of LDA. A related question has to do with improving upon a linear classifier’s weight vector, which cannot be tuned or corrected in the same way. Relying on the intuition that a good weight vector should be able to extract the maximum discriminatory information content from the data point being classified, we show in this work that tuning the multidimensional weight vector can indeed be reduced to tuning a scalar. The next section details our contributions in this regard.

## 5.2 Contributions

In the first part of this work, we show that any binary linear classifier discriminant can be decomposed into terms containing discriminating information and non-discriminating noise. A linear form of this decomposition parameterized by a variable  $\alpha$  controls the trade-off between conflicting noise and information terms. At the optimal setting of  $\alpha$ , the modified discriminant performs at least as good as the original classifier from which it was produced. Following this, the effect of the weight vector modification on the performance of an assortment of linear

classifiers under different data dimensionality and sample size scenarios is studied. The method specifically yields significant performance gains for the LDA classifier under high estimation noise. Interestingly, the parameterized LDA operates as a bridge between LDA and the nearest centroid classifier, and performs at least as good as either of these classifiers. This is an example of a bias-variance tradeoff as the nearest centroid discriminant assumes isotropic covariance (which introduces bias) and the LDA discriminant employs the sample covariance estimator based on the data (which introduces variance). Additionally, it is shown that tuning the weight vector according to the proposed method can significantly improve the performance of certain classifiers whose native hyperparameters are not optimally set. It is shown that with weight vector tuning, the SVM with non-optimally tuned penalty can achieve performance close to that of its tuned counterpart. In this case, tuning the weight vector is fundamentally different from tuning the native hyperparameter of the classifier as it occurs post weight vector generation, while the native hyperparameter tuning occurs prior to weight vector generation. For SVM, generating the weight vector for each value of the native hyperparameter involves solving an optimization problem. Tuning the weight vector according to the proposed method, however, reduces to a simple grid search over a scalar parameter. This idea can be generalized to any classifier with hyperparameters that are set prior to weight vector generation.

The remainder of this work consists of an asymptotic study of the parameterized LDA classifier under a growth regime in which the data dimensionality and sample size grow proportionally. We use random matrix theory to show that the probability of misclassification of this classifier converges to a limit that is a function of the true class statistics. We also derive a consistent estimator of the probability of misclassification by which the classifier parameter  $\alpha$  can be tuned. This estimator is more computationally efficient than other tuning methods which rely on additional testing points or recycling the training set, e.g. CV, as it requires no additional testing points and no averaging. We demonstrate its

performance on real data.

An additional finding of this work is a new interpretation of the optimality of LDA. As shown in Chapter 2, the LDA decision rule is optimal in the Fisher's linear discriminant sense, the least squares sense, and also - since it is the estimated Bayes rule - in the posterior probability sense (asymptotically when  $n \rightarrow \infty$  for fixed  $p$ ). Moreover, we show in this chapter that the Bayes weight vector,  $\mathbf{w}_{\text{Bayes}} := \Sigma^{-1}\boldsymbol{\mu}$  - which LDA estimates - is optimal in the sense that it achieves the minimum noise (in the mean square error sense) with respect to the test point when the classes are Gaussian with common covariance.

To summarize, the main contributions of this work are

- A practical method for weight vector tuning which reduces to grid search over a scalar parameter
- A novel interpretation of the optimality of the LDA classifier in terms of minimizing test point noise
- Asymptotic expressions for the probability of misclassification of the parameterized LDA classifier
- A consistent estimator of the probability of misclassification of the parameterized LDA classifier

We now present our proposed weight vector tuning procedure.

### 5.3 Weight vector tuning procedure

Consider the linear classifier defined in Chapter 2 with discriminant of the form (2.2) and corresponding classifier of the form (2.1). Examples of classifiers which fit this form include LDA, SVM and least-squares SVM (both using linear kernels), and R-LDA. In this work, we propose a method of tuning the weight vector  $\mathbf{w}$ , which reduces the non-discriminative 'noisy' components of the original discriminant (2.2). As a result, the modified discriminant achieves a testing error rate at least as good as the original and, in certain cases, much better.

In this particular chapter, unlike the rest of this dissertation, we generalize the common covariance assumption to allow for distinct covariances. More specifically, let the means and covariances of classes  $\mathcal{C}_0$  and  $\mathcal{C}_1$  be denoted by  $\boldsymbol{\mu}_0$ ,  $\boldsymbol{\Sigma}_0$  and  $\boldsymbol{\mu}_1$ ,  $\boldsymbol{\Sigma}_1$  respectively. In Section 5.3.1, we explore an ideal case in which the discriminant neatly decomposes into separate information and noise terms and the noises cancel out optimally in a linear fashion under the assumption of perfectly known means and that  $\mathcal{C}_0$  and  $\mathcal{C}_1$  makeup a Gaussian mixture model with common class covariance  $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$ . Inspired by the findings of Section 5.3.1, in Section 5.3.2 we heuristically extend this result to a more practical scenario which assumes unknown means and no restriction on the class distributions.

### 5.3.1 Known class means

In this section, assume that the data distribution means  $\boldsymbol{\mu}_0$  and  $\boldsymbol{\mu}_1$  are known exactly and that  $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$ . We proceed to derive a noise-minimized version of (2.2).

Consider the shifted test point  $\tilde{\mathbf{x}} = \mathbf{x} - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2}$ . For any given classifier with weight vector  $\mathbf{w}$ , we show that the projection of  $\tilde{\mathbf{x}}$  onto  $\mathbf{w}$ , i.e.,  $\mathbf{w}^T \tilde{\mathbf{x}}$ , can be decomposed into ‘informative’ components which aid in discriminating the class of  $\mathbf{x}$  and ‘noisy’ components which interfere with discriminating the class of  $\mathbf{x}$ . We then take advantage of this hidden structure for the purpose of reducing the overall noise and obtaining a better classifier.

Recall from Section 2.3.1 that  $\boldsymbol{\mu} := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ . The expression  $\tilde{\mathbf{x}}$  can be expressed as the sum of its projection onto  $\boldsymbol{\mu}$  and projection orthogonal to  $\boldsymbol{\mu}$  as

$$\tilde{\mathbf{x}} = \frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{\boldsymbol{\mu}^T\boldsymbol{\mu}}\tilde{\mathbf{x}} + \mathbf{P}_\mu\tilde{\mathbf{x}} \quad (5.1)$$

where  $\mathbf{P}_\mu = \left(\mathbf{I} - \frac{\boldsymbol{\mu}\boldsymbol{\mu}^T}{\boldsymbol{\mu}^T\boldsymbol{\mu}}\right)$  is the projection orthogonal to  $\boldsymbol{\mu}$ . Substituting (5.1) into

$\mathbf{w}^T \tilde{\mathbf{x}}$  results in the decomposition of  $\mathbf{w}^T \tilde{\mathbf{x}}$  as

$$\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \tilde{\mathbf{x}} + \mathbf{w}^T \mathbf{P}_\mu \tilde{\mathbf{x}} \quad (5.2)$$

We now show that the first term in (5.2) is composed of an informative component and noisy component with respect to  $\mathbf{x}$ , while the second term consists solely of noise. Assume  $\mathbf{x} \in \mathcal{C}_i$ , where  $i$  is either 0 or 1. Then, assuming the Gaussian mixture model (2.3), we have  $\mathbf{x} | \mathbf{x} \in \mathcal{C}_i \sim \boldsymbol{\mu}_i + \boldsymbol{\Sigma}^{1/2} \mathbf{z}$ ,  $i = 0, 1$ , where  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . The first term in (5.2), conditioned on the class of  $\mathbf{x}$ , is then distributed as follows

$$\begin{aligned} \frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \tilde{\mathbf{x}} | \mathbf{x} \in \mathcal{C}_i &\sim \frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \left( (-1)^{i+1} \frac{\boldsymbol{\mu}}{2} + \boldsymbol{\Sigma}^{1/2} \mathbf{z} \right) \\ &= \underbrace{(-1)^{i+1} \frac{\mathbf{w}^T \boldsymbol{\mu}}{2}}_{I_1(\text{information})} + \underbrace{\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{1/2} \mathbf{z}}_{N_1(\text{noise})} \end{aligned} \quad (5.3)$$

The first term in (5.3) carries information about the class of  $\mathbf{x}$  through its sign. The second term is the same regardless of the class of  $\mathbf{x}$  and therefore carries no discriminating information. This is a direct result of assuming a common covariance between  $\mathcal{C}_0$  and  $\mathcal{C}_1$ . The informative component is denoted by  $I_1$  while the noisy component with respect to  $\mathbf{x}$  is denoted by  $N_1$ . Similarly,

$$\begin{aligned} \mathbf{w}^T \mathbf{P}_\mu \tilde{\mathbf{x}} | \mathbf{x} \in \mathcal{C}_i &\sim \mathbf{w}^T \mathbf{P}_\mu \left( (-1)^{i+1} \frac{\boldsymbol{\mu}}{2} + \boldsymbol{\Sigma}^{1/2} \mathbf{z} \right) \\ &= \underbrace{\mathbf{w}^T \mathbf{P}_\mu \boldsymbol{\Sigma}^{1/2} \mathbf{z}}_{N_2(\text{noise})} \end{aligned}$$

The discriminatory component of this term is lost in the orthogonal projection, and therefore this term consists solely of noise with respect to the testing point, denoted by  $N_2$ .

To recap, the decomposition of the weight vector divides the discriminant into a single observable term containing  $I_1$  and  $N_1$  and a single observable term containing  $N_2$ . Without the decomposition, none of these individual noise/infor-

mation terms are accessible. Now, in the interest of achieving better classification performance, we wish to reduce the overall noise content in the discriminant. We can leverage the observable term containing  $N_2$  to bring out the information in the observable term containing both information  $I_1$  and noise  $N_1$ . To this end, consider the following modification of the discriminant (5.2),

$$\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \tilde{\mathbf{x}} + g(\mathbf{w}^T \mathbf{P}_\mu \tilde{\mathbf{x}}) \quad (5.4)$$

for any function  $g(\cdot)$ , and which, by the above analysis, is equivalent to

$$I_1 + N_1 + g(N_2)$$

The optimal  $g(\cdot)$  such that

$$\mathbb{E}[(N_1 + g(N_2))^2]$$

is minimized is the Minimum Mean Square Error (MMSE) estimator  $\mathbb{E}[-N_1|N_2]$ . This choice of  $g(\cdot)$  has the effect of minimizing the total noise in the discriminant in the mean square error sense. We show in Section 5.3.1 that it simultaneously minimizes the probability of misclassification. In the following Lemma 1, we derive the exact form of  $g(\cdot)$  for a given  $\mathbf{w}$  based on the class distribution assumptions (2.3).

**Lemma 1** *The optimal  $g(N_2)$  is the linear function of  $N_2$  given by  $g^*(N_2) = \alpha_{\text{MMSE}}(\mathbf{w})N_2$ , where*

$$\alpha_{\text{MMSE}}(\mathbf{w}) = -\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \frac{\boldsymbol{\mu}^T \boldsymbol{\Sigma} \mathbf{P}_\mu \mathbf{w}}{\mathbf{w}^T \mathbf{P}_\mu \boldsymbol{\Sigma} \mathbf{P}_\mu \mathbf{w}}. \quad (5.5)$$

**Proof:** Given  $\mathbf{w}$ ,

$$-N_1 = -\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{1/2} \mathbf{z} \sim \mathcal{N}\left(0, \left(\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}}\right)^2 \boldsymbol{\mu}^T \boldsymbol{\Sigma} \boldsymbol{\mu}\right)$$

and

$$N_2 = \mathbf{w}^T \mathbf{P}_\mu \Sigma^{1/2} \mathbf{z} \sim \mathcal{N}(0, \mathbf{w}^T \mathbf{P}_\mu \Sigma \mathbf{P}_\mu \mathbf{w})$$

are jointly Gaussian random variables. Thus, the optimal  $g^*(N_2) = \mathbb{E}[-N_1|N_2]$  reduces to a linear function of  $N_2$  given by

$$\begin{aligned} g^*(N_2) &= \frac{\text{Cov}[-N_1, N_2]}{\text{Var}[N_2]} (N_2 - \mathbb{E}[N_2]) + \mathbb{E}[-N_1] \\ &= -\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \frac{\boldsymbol{\mu}^T \Sigma \mathbf{P}_\mu \mathbf{w}}{\mathbf{w}^T \mathbf{P}_\mu \Sigma \mathbf{P}_\mu \mathbf{w}} N_2. \end{aligned}$$

Note that  $N_2$  is observable only through the expression  $\mathbf{w}^T \mathbf{P}_\mu \tilde{\mathbf{x}}$  and so when using this result we replace  $N_2$  by its observable counterpart.

Based on this result, we have the following theorem.

**Theorem 1** *The discriminant that minimizes the noise with respect to the test point in the Mean Square Error (MSE) sense for a given  $\mathbf{w}$ , known means, and under the data distribution assumptions of (2.3), is*

$$\frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \tilde{\mathbf{x}} + \alpha_{\text{MMSE}}(\mathbf{w}) \mathbf{w}^T \mathbf{P}_\mu \tilde{\mathbf{x}}, \quad (5.6)$$

or, equivalently,

$$\mathbf{w}'^T \mathbf{x} + w'_0,$$

where

$$\mathbf{w}' = \frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu} + \alpha_{\text{MMSE}}(\mathbf{w}) \mathbf{P}_\mu \mathbf{w}$$

and

$$w'_0 = -\frac{1}{2} \left( \frac{\mathbf{w}^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu} + \alpha_{\text{MMSE}}(\mathbf{w}) \mathbf{P}_\mu \mathbf{w} \right)^T (\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1).$$

This result is obtained by simply evaluating (5.4) using  $g^*(\cdot)$ . We make several remarks concerning this result. Firstly, the modified discriminant is linear. This is a direct result of the Gaussian assumption (2.3), which, while not technically necessary, is desirable, as it produces a simple linear form which inspires the parameterized formulation presented in the next section. Secondly, the original

weight vector  $\mathbf{w}$  is modified to  $\mathbf{w}'$  and a bias  $w'_0$  is generated. This bias is the optimal bias in the sense of minimizing the probability of misclassification under the class distribution assumptions of (2.3) and equal class priors when fixing the weight vector to  $\mathbf{w}'$  (see Mai et al. (2012) Proposition 2). Finally, viewing the modified discriminant (5.6) as a function of a parameter  $\alpha$  as follows

$$\frac{\mathbf{w}'^T \boldsymbol{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\mu}} \boldsymbol{\mu}^T \tilde{\mathbf{x}} + \alpha \mathbf{w}'^T \mathbf{P}_\mu \tilde{\mathbf{x}}, \quad (5.7)$$

$\alpha = \alpha_{\text{MMSE}}(\mathbf{w}')$  yields a stationary point of its probability of misclassification and achieves the minimum probability of misclassification when  $\mathbf{w}'^T \boldsymbol{\mu} > 0$ . This is demonstrated in Section 5.3.1.

The following corollary of Theorem 1 lends intuition as well as credibility to this technique by showing that it recovers the Bayes optimal classifier discriminant for the assumed class distributions from its weight vector. The Bayes classifier in this case is linear. It is the (known statistics) LDA classifier, with discriminant

$$\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \tilde{\mathbf{x}} + \ln \frac{\pi_1}{\pi_0}. \quad (5.8)$$

To avoid confusion with the estimated LDA classifier used in practice, we refer to (5.8) as the ‘Bayes’ classifier whose corresponding weight vector is  $\mathbf{w}_{\text{Bayes}} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ .

**Corollary 1** *Computing the parameter (5.5) corresponding to the Bayes classifier (5.8) yields*

$$\alpha_{\text{MMSE}}(\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}) = 1$$

*and the resulting discriminant (5.6) recovers the Bayes discriminant in (5.8) when the class priors are equal.*

Since there is no modification of the weight vector, we conclude that the Bayes weight vector (in the case of known statistics) is optimal relative to itself in that it achieves the minimum noise (in the MSE sense) with respect to the test point

under the assumed class distributions.

## Experiments with known means

For the following simulation and any simulations involving synthetic data in the remainder of this paper, the exact expected testing error/probability of misclassification of a linear classifier learned on a given training set is computed using knowledge of the data distribution from which the testing data is generated. All synthetic data in this chapter is generated from a two-class Gaussian mixture model. The exact testing error for a given training set is computed using (2.4).

Now consider the parameterized version (5.7) of (5.6). The objective of the following simulation is to show that  $\alpha_{\text{MMSE}}(\mathbf{w})$  given by (5.5) coincides with the  $\alpha$  yielding a stationary point of the expected testing error of (5.7). The stationary point is a minimum when  $\mathbf{w}^T \boldsymbol{\mu} > 0$  and is otherwise a maximum, as in that case, the orientation of  $\mathbf{w}$  flips the class labels.

To demonstrate this, a weight vector  $\mathbf{w}$  is uniformly sampled from all  $\mathbf{w}$  such that  $\|\mathbf{w}\|_2 = 1$  using the method in (Weisstein, 2017). It is then fed to (5.7) and the exact expected testing error with varying  $\alpha$  is plotted using (2.4). The quantity  $\alpha_{\text{MMSE}}(\mathbf{w})$  is then computed from (5.5) for comparison. The class statistics used for this simulation are

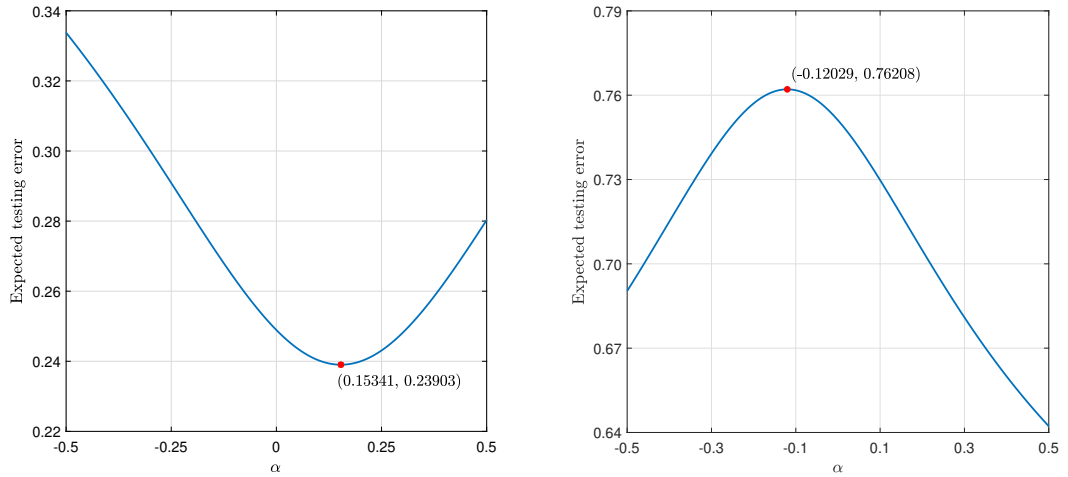
$$\boldsymbol{\mu}_0 = \frac{1}{p^{1/4}} \left[ \mathbf{1}_{\lceil \sqrt{p} \rceil}^T \quad \mathbf{0}_{p - \lceil \sqrt{p} \rceil - 2}^T \quad 2 \quad 2 \right]^T, \quad \boldsymbol{\mu}_1 = \mathbf{0}_p, \quad (5.9)$$

and

$$\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \frac{10}{p} \mathbf{1}_p \mathbf{1}_p^T + 0.1 \mathbf{I}_p \quad (5.10)$$

where  $p = 200$ . Here,  $\pi_0 = \pi_1 = 0.5$ .

Figure 5.1a and Figure 5.1b show the results when  $\mathbf{w}^T \boldsymbol{\mu} > 0$  and  $\mathbf{w}^T \boldsymbol{\mu} < 0$ , respectively. In Figure 5.1a, the minimum expected testing error occurs at  $\alpha = 0.15341$ . This exactly coincides with  $\alpha_{\text{MMSE}}(\mathbf{w})$  of Theorem 1 that minimizes the noise in the discriminant. In Figure 5.1b, the *maximum* expected testing



(a) Here  $\mathbf{w}^T \boldsymbol{\mu} > 0$  and  $\alpha_{\text{MMSE}}(\mathbf{w}) = 0.1534$  coincides with the  $\alpha$  yielding the minimum expected testing error  
 (b) Here  $\mathbf{w}^T \boldsymbol{\mu} < 0$  and  $\alpha_{\text{MMSE}}(\mathbf{w}) = -0.1203$  coincides with the  $\alpha$  yielding the maximum expected testing error

Figure 5.1: Plot of the expected testing error of the modified discriminant against  $\alpha$  for a randomly generated weight vector  $\mathbf{w}$

error occurs at  $\alpha = -0.12029$ , which, again, exactly coincides with  $\alpha_{\text{MMSE}}(\mathbf{w})$  that minimizes the noise in the discriminant. The latter discriminant's behavior can be explained by the fact that the orientation of the randomly generated  $\mathbf{w}$  flips the class labels. Simply taking the negative of  $\mathbf{w}$  yields a classifier having the *minimum* expected testing error at  $\alpha_{\text{MMSE}}(\mathbf{w})$ . In conclusion, minimizing the noise in the discriminant in the MSE sense is equivalent to minimizing the expected testing error, as long as  $\mathbf{w}$  is sensibly oriented. This motivates using this criteria as the basis for designing a better classifier in the next section.

### 5.3.2 Unknown class means

The previous section derives the discriminant with minimum noise with respect to the test point for a general binary linear classifier with weight vector  $\mathbf{w}$  under the assumption of Gaussian classes with known means and a common covariance. A more practical scenario is when all class statistics are unknown and sample statistics are used instead. Using the sample mean estimates (defined in Section 2.3.2) introduces an additional estimation noise into the discriminant.

Let  $\hat{\mathbf{x}} = \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}$  where  $\hat{\boldsymbol{\mu}} := \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ , as defined in Section 2.3.2. Given a

weight vector,  $\mathbf{w}$ ,  $\mathbf{w}^T \hat{\mathbf{x}}$  can be expressed as

$$\mathbf{w}^T \hat{\mathbf{x}} = \frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \hat{\mathbf{x}} + \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \hat{\mathbf{x}} \quad (5.11)$$

where  $\mathbf{P}_{\hat{\boldsymbol{\mu}}} = \left( \mathbf{I} - \frac{\hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \right)$ . Regardless of the class distributions and whether assuming distinct covariances  $\boldsymbol{\Sigma}_0$  and  $\boldsymbol{\Sigma}_1$  or common class covariances  $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$ , following a similar line of logic to the analysis in Section 5.3.1 reveals that, while the first term in (5.11) is similarly composed of both information and noise (whether that be estimation noise, noise from the test point, or both), the second term is not purely noise. In fact, it is informative. This is shown in detail in Appendix C.1.

Thus, when the means are unknown, the approach taken in Section 5.3.1 of minimizing the squared sum of ‘noise 1’ with the second term no longer applies, as the second term is informative. Nonetheless, the interaction of this term with the noise in the first term can potentially yield performance gains and so motivated by Section 5.3.1, the following parameterized version of the sample statistic equivalent of (5.6) is proposed

$$\frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \hat{\mathbf{x}} + \alpha \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \hat{\mathbf{x}} \quad (5.12)$$

where  $\alpha$  is a parameter to be tuned.

The following Section 5.3.2 demonstrates that a better misclassification rate may be achieved by setting  $\alpha$  to a value that is not equal to one (where  $\alpha = 1$  recovers the original projection with optimal bias assuming equal priors and the class distribution in (2.3)). A significant improvement is observed when the estimation noise is high.

## Experiments with Unknown Means

In this section we explore the behavior of (5.12) under a variety of settings and for an assortment of starting weight vectors. We first list and briefly describe

the discriminants from which these weight vectors are extracted, namely, LDA, logistic regression, linear SVM, R-LDA, and the discriminant-averaging RP-LDA ensemble classifier.

- As discussed in Chapter 2, **LDA** in the form (5.8) is the Bayes classifier for data distributed as (2.3). In practice, the class statistics are unknown and sample estimates are used instead. Its weight vector in this case is  $\mathbf{w}_{\text{LDA}} = \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$ .
- For linearly separable training data, **SVM with linear kernel** (see (Friedman et al., 2001)) finds a hyperplane that maximizes the margin between one class and the other subject to constraints of perfect classification on the training points. When the training data is linearly inseparable, the constraints are relaxed by penalizing each (possibly) misclassified point. The penalty is a parameter that must be tuned. This variant is called the soft-margin SVM with linear kernel, and it is what we use in this paper.
- **Logistic regression** (see (Friedman et al., 2001)) models the log-odds  $\ln \left( \frac{P[\mathbf{x} \in \mathcal{C}_1 | \mathbf{x}]}{1 - P[\mathbf{x} \in \mathcal{C}_1 | \mathbf{x}]} \right)$  as a linear function of the test point. The decision boundary corresponds to the set of points at which the log-odds equals zero. The weight vector and bias of the decision boundary are learned by maximizing the likelihood of the training data.
- **R-LDA** counters the small sample issue in LDA by regularizing the pooled sample covariance estimate before inverting it. There are several possibilities for the form of the regularization (see (Guo et al., 2007)). Here, we opt for

$$\hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \gamma \mathbf{I}_p \right)^{-1} \hat{\mathbf{x}} + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0},$$

where  $\gamma$  is the regularization parameter that must be tuned. The weight vector here is  $\mathbf{w}_{\text{R-LDA}} = \left( \hat{\Sigma} + \gamma \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}}$ .

- As explained in Chapter 3, the **discriminant-averaging RP-LDA en-**

**semble** (see (Durrant and Kabán, 2013)) counters the small sample issue in LDA by reducing the dimensionality of the training samples (and test point) using random matrices. Each projection  $\mathbf{R}_k \in \mathbb{R}^{d \times p}$  yields a discriminant. These are averaged over  $M$  projections so that the final discriminant has the form (3.4). The weight vector is  $\mathbf{w}_{\text{disc-avg}} = \frac{1}{M} \sum_{k=1}^M \mathbf{R}_k^T (\mathbf{R}_k \hat{\Sigma} \mathbf{R}_k^T)^{-1} \mathbf{R}_k \hat{\boldsymbol{\mu}}$ . The reduced dimension  $d$  is a parameter that must be tuned.

For these simulations, we consider two data distributions: data generated from classes having a common covariance and data generated from classes having distinct covariance matrices. We also consider three regimes of  $n$  versus  $p$ :  $n$  on the order of  $p$  ( $p = 400, n = 450$ ),  $n > p$  ( $p = 10, n = 500$ ), and  $n < p$  ( $p = 300, n = 100$ ). We apply the appropriate classifiers to each regime. LDA requires  $n > p$ , soft-margin SVM is applicable in any regime, logistic regression requires  $n$  be much greater than  $p$  to ensure convergence of the maximum likelihood estimates of the weight vector and bias, and finally, R-LDA and the discriminant-averaging RP-LDA ensemble are designed for the regime  $n < p$ .

Each classifier is trained on a generated training set. Additionally, for SVM, R-LDA, and RP-LDA, the penalty,  $\gamma$ , and  $d$  parameters are chosen to minimize the expected testing error given that training set. The SVM penalty is tuned within the set  $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$ ,  $\gamma$  within the set  $[10^{-4}, 2]$ , in increments of 0.1, and  $d$  from 1 to the maximum allowable setting of  $d = \text{rank}(\hat{\Sigma}) - 2$ , in increments of 2. After this is done, we have a weight vector  $\mathbf{w}$  for each classifier. Each weight vector is fed into (5.12) to obtain an  $\alpha$ -parameterized version of the discriminant. Let us refer to these new classifiers as  $\alpha$ -LDA,  $\alpha$ -SVM,  $\alpha$ -log,  $\alpha$ -RLDA, and  $\alpha$ -RPLDA for short. For each  $\alpha$ -parameterized discriminant, we vary  $\alpha$  and compute the expected testing error using (2.4). These errors are averaged over 100 independently generated training sets. Error bars depicting the standard errors are plotted alongside this average.

Recall that setting  $\alpha = 1$  in (5.12) produces a discriminant having the original weight vector  $\mathbf{w}$  and a bias with minimum probability of misclassification (under

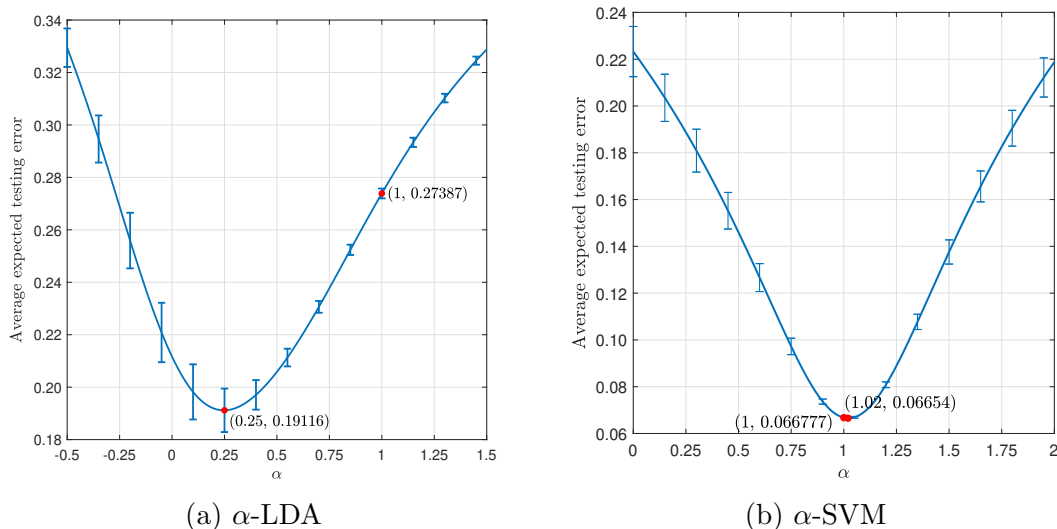


Figure 5.2: Plots of expected testing error averaged over 100 training sets for data generated from classes with a common  $\Sigma$ . Here,  $p = 400$  and  $n = 450$ .

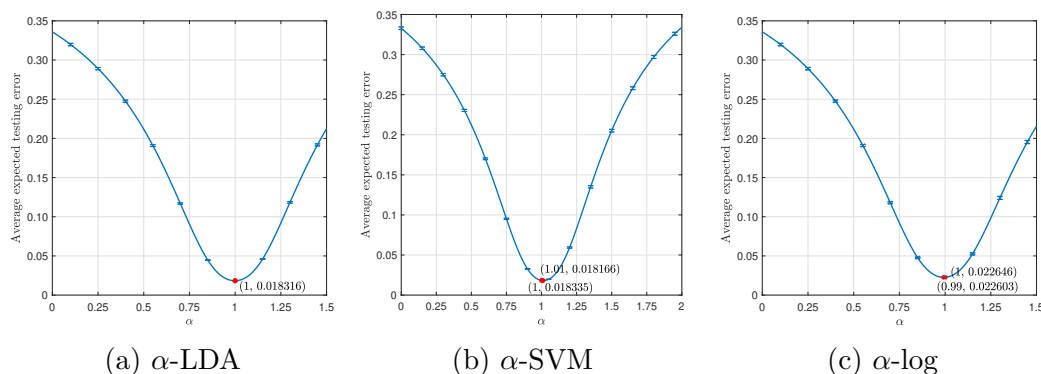


Figure 5.3: Plots of expected testing error averaged over 100 training sets for data generated from classes with a common  $\Sigma$ . Here,  $p = 10$  and  $n = 500$ .

the Gaussian mixture model and equal priors assumption) for that weight vector. In what follows, we use  $\alpha = 1$  as a reference point for determining whether or not there is a significant improvement in classifier performance at the  $\alpha$  achieving the minimum error rate. To quantify the improvement, we report percentage changes relative to the average expected testing error at  $\alpha = 1$  computed as  $\frac{\text{error at } \alpha \text{ achieving the minimum} - \text{error at } \alpha=1}{\text{error at } \alpha=1} \times 100$ . This quantity reflects the fact that a given error improvement starting at an already low error rate at the baseline  $\alpha = 1$  is more significant than when the error is high to start with.

The first set of class statistics we consider are (5.9), (5.10), and  $\pi_0 = \pi_1 = 0.5$ . Corresponding to this data distribution are Figures 5.2, 3.3, and 5.4.

Figures 5.2a and 5.2b plot the average expected testing errors of  $\alpha$ -LDA and  $\alpha$ -

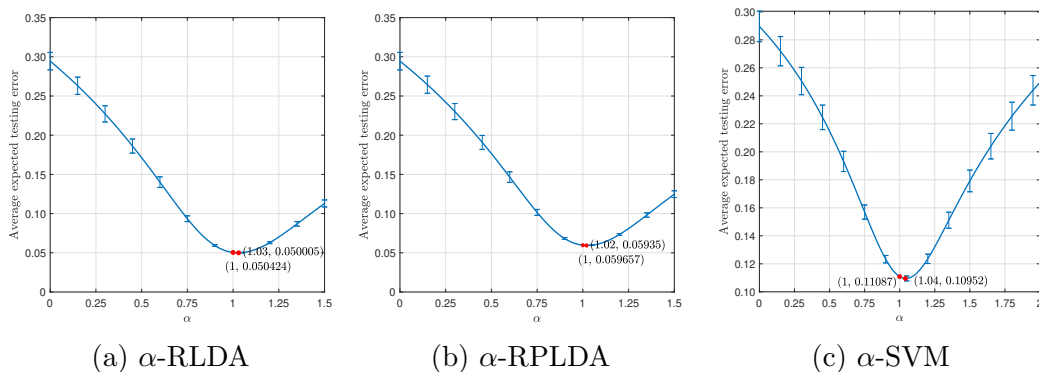


Figure 5.4: Plots of expected testing error averaged over 100 training sets for data generated from classes with a common  $\Sigma$ . Here,  $p = 300$  and  $n = 100$ .

SVM respectively against varying  $\alpha$  when  $p = 400$  and  $n = 450$ . At  $\alpha = 0.25$ , the  $\alpha$ -LDA classifier achieves a 30.2% relative decrease in the average expected testing error. Note that ordinary LDA ( $\alpha = 1$ ) is nowhere near optimal. On the other hand,  $\alpha$ -SVM achieves a 0.355% decrease in average expected testing error at  $\alpha = 1.02$ . These results suggest that there is a lot to be gained performance-wise by LDA in this regime but not so much by linear SVM. This can be attributed to the fact that LDA relies on sample estimation and that the noise due to estimation is high when  $p = 400$  and  $n = 450$ . This is further supported by the results of Figures 5.3a, 5.3b and 5.3c, which plot the average expected testing errors of  $\alpha$ -LDA,  $\alpha$ -SVM, and  $\alpha$ -log, respectively against varying  $\alpha$  when  $p = 10$  and  $n = 500$ . The minimum average expected occurs at exactly  $\alpha = 1$  for  $\alpha$ -LDA,  $\alpha = 1.01$  for  $\alpha$ -SVM and at  $\alpha = 0.99$  for  $\alpha$ -log, with the latter two classifiers achieving a relative decrease of no more than 1% and 0.2% respectively. The extreme behavior in all three figures can be explained by the fact that there is very little estimation noise for this choice of dimensions. What is notable is the difference between Figure 5.2a and Figure 5.3a which suggests that the weight vector tuning method is most effective under high estimation noise and for methods which are most sensitive to it. This idea is again reinforced in Figures 5.4a, 5.4b, and 5.4c, in which the average expected testing errors of  $\alpha$ -RLDA,  $\alpha$ -RPLDA, and  $\alpha$ -SVM respectively are plotted against varying  $\alpha$  when  $p = 300$  and  $n = 100$ . The relative decrease in errors for each of the three classifiers does

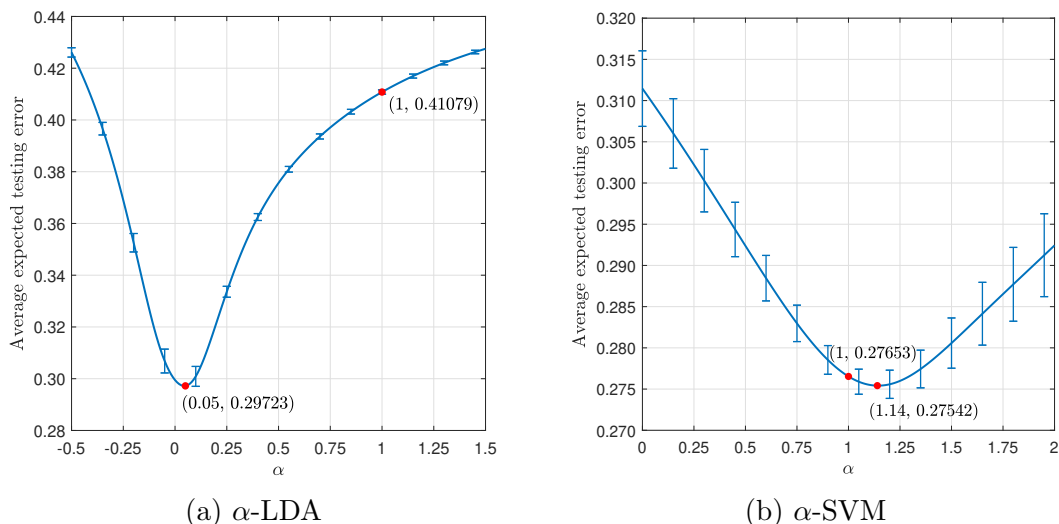


Figure 5.5: Plots of expected testing error averaged over 100 training sets for data generated from classes with distinct  $\Sigma_0$  and  $\Sigma_1$ . Here,  $p = 400$  and  $n = 450$ .

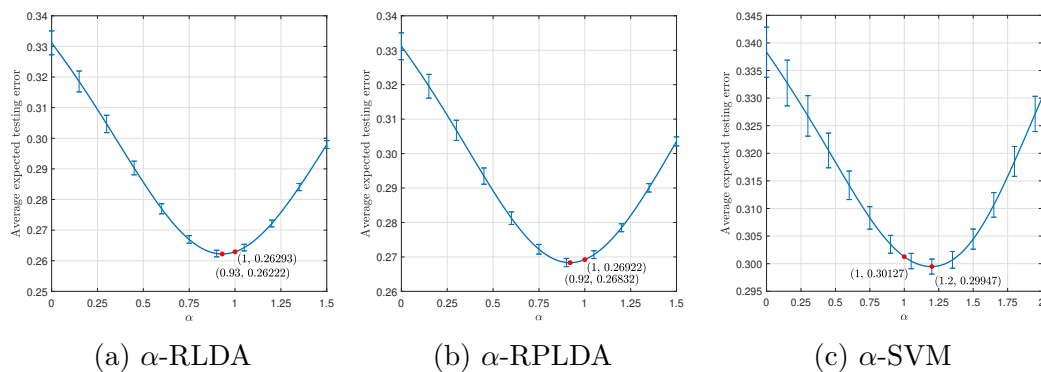


Figure 5.6: Plots of expected testing error averaged over 100 training sets for data generated from classes with distinct  $\Sigma_0$  and  $\Sigma_1$ . Here,  $p = 300$  and  $n = 100$ .

not exceed 1.3%. It must be that R-LDA and RP-LDA are able to reduce much of the estimation noise on their own, and so the  $\alpha$  parameterization does not bring much improvement.

Figures 5.5 and 5.6 are based on data with the class statistics

$$\mu_0 = \frac{1}{p^{1/4}} \left[ \mathbf{1}_{\lceil \sqrt{p} \rceil}^T \quad \mathbf{0}_{p - \lceil \sqrt{p} \rceil - 2}^T \quad 2 \quad 2 \right]^T, \quad \mu_1 = \mathbf{0}_p,$$

$$[\Sigma_0]_{ij} = 0.9^{|i-j|}, \quad i, j = 1, \dots, p,$$

$$\Sigma_1 = \frac{10}{p} \mathbf{1}_p \mathbf{1}_p^T + 0.1 \mathbf{I}_p,$$

and  $\pi_0 = \pi_1 = 0.5$ . The difference here is that the class covariances are distinct.

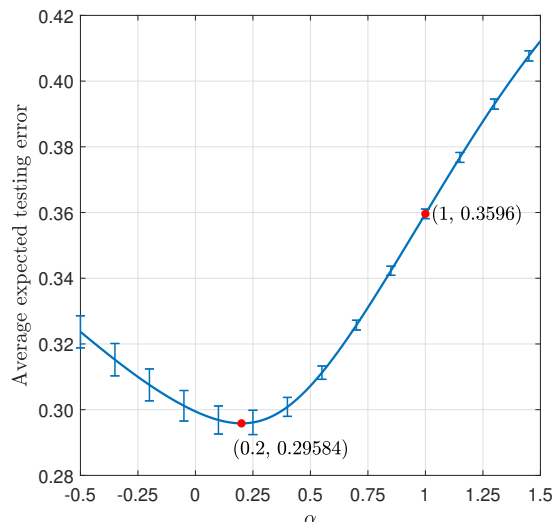


Figure 5.7: Plot of expected testing error of  $\alpha$ -SVM with penalty set to 1 averaged over 100 training sets for data generated from classes with distinct  $\Sigma_0$  and  $\Sigma_1$ . Here,  $p = 400$  and  $n = 450$ .

Figures 5.5a and 5.5b again plot the average expected testing errors of  $\alpha$ -LDA and  $\alpha$ -SVM, respectively, against varying  $\alpha$  when  $p = 400$  and  $n = 450$ . In this case,  $\alpha$ -LDA significantly improves in performance when  $\alpha$  is set to a non-unit value. It achieves a relative decrease in error of 27.6% at  $\alpha = 0.05$ , while  $\alpha$ -SVM achieves a relative decrease in error of 0.4% at  $\alpha = 1.14$ . Finally, Figures 5.6a, 5.6b, and 5.6c plot the average expected testing errors of  $\alpha$ -RLDA,  $\alpha$ -RPLDA and  $\alpha$ -SVM against varying  $\alpha$  when  $p = 300$  and  $n = 100$ . Here, the relative decreases in error do not exceed 0.6%.

As described at the beginning of this section, for each training set, the SVM penalty is tuned to the value yielding the lowest expected testing error. We found that SVM does not show much improvement when it is  $\alpha$  parameterized. It is interesting to observe what happens when the penalty is not tuned beforehand. Instead we set the penalty to 1 (its default setting in the MATLAB R2019b ‘fitsvm’ function) uniformly across all training sets. Figure 5.7 shows the resulting average expected testing error of  $\alpha$ -SVM plotted against vary  $\alpha$  in the same setting as in Figure 5.5b, i.e.  $p = 450$ ,  $n = 400$ , and distinct  $\Sigma_0$  and  $\Sigma_1$ . In this case,  $\alpha$ -SVM achieves a relative decrease in error of 17.7% at  $\alpha = 0.2$ . Clearly, the method improves performance when  $\mathbf{w}$  itself is not at its optimal.

Taking this idea further, we show that tuning the weight vector of a SVM classifier with a poorly chosen penalty can compensate for the resulting loss in performance. Figure 5.8 is based on the USPS dataset consisting of separate training and testing sets of grayscale images of handwritten digits 0 – 9. Pairs of digits are used to form a binary classification problem. For each pair of digits, a poorly-tuned SVM classifier is  $\alpha$  parameterized and the testing error plotted against  $\alpha$  to illustrate the effect of weight vector tuning.

For the digit pair ‘2’ and ‘6’, an optimized SVM classifier can achieve a testing error of 0.0217. Figure 5.8a shows the testing error of  $\alpha$ -SVM starting with a poorly-tuned SVM classifier whose testing error on this digit pair is 0.0489. By weight vector tuning, the testing error can be brought down to 0.0272. This is comparable to the performance of the original optimized SVM classifier. Similarly, for the digit pair ‘3’ and ‘5’, an optimized SVM classifier can achieve a testing error of 0.0675. Figure 5.8a shows the testing error of  $\alpha$ -SVM starting with a poorly-tuned SVM classifier whose testing error on this digit pair is 0.0951. By weight vector tuning, the testing error can be brought down to 0.0736.

The significance of this finding is the potential savings in computation that can be made by weight vector tuning versus penalty tuning. The reason for this is that weight vector tuning is an afterthought; it occurs post weight vector generation. On the other hand, setting the penalty is done prior to weight vector generation. An optimization problem must be solved to generate the weight vector with each setting of the penalty. At best, generating this weight vector has a complexity of  $\mathcal{O}(n^2)$  at each setting of the penalty (Bottou and Lin, 2007).

This idea generalizes to any linear classifier whose native hyperparameters are set prior to weight vector generation. The tuning of the hyperparameters will then involve repeatedly generating the weight vector. If this process is costly, weight vector tuning can provide a more computationally efficient method of improving performance than tuning the native hyperparameters. Another example that is not demonstrated here is the RP-LDA ensemble classifier whose projection

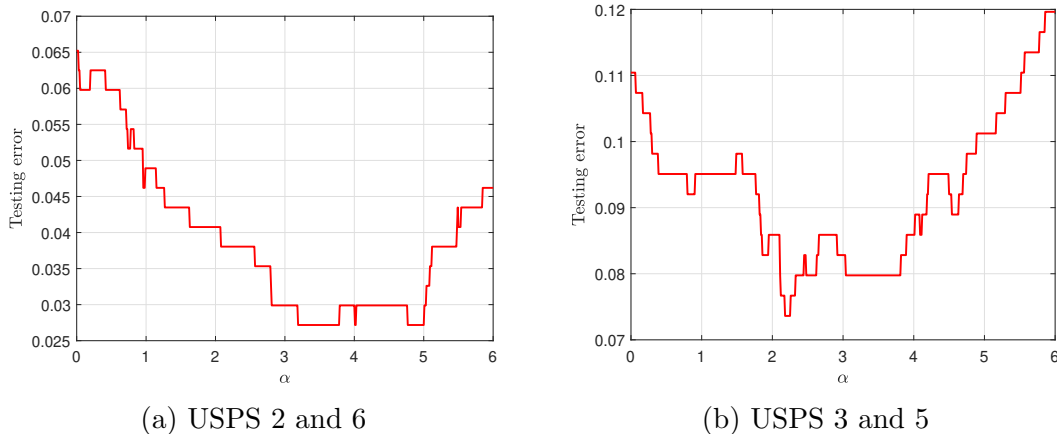


Figure 5.8: Plots of testing error on USPS digit pairs of  $\alpha$ -SVM with penalty set non-optimally

dimension  $d$  is a native hyperparameter. Tuning this is computationally inefficient as it means projecting all the data with each setting of  $d$ . A simple alternative is weight vector tuning.

Overall, we conclude from this section that  $\alpha$ -LDA in the ‘ $n$  on the order of  $p$ ’ scenario shows the most promise in terms of improved performance. For this reason, we proceed to study this classifier in the RMT asymptotic regime in the next section.

## 5.4 Asymptotic analysis and tuning of the parameterized LDA classifier

In this section, we extend our study of  $\alpha$ -LDA, the modified weight discriminant (5.6) corresponding to the plugin LDA weight vector. The  $\alpha$ -LDA discriminant

$$\frac{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \hat{\mathbf{x}} + \alpha \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \hat{\mathbf{x}}$$

is a bridging between LDA (when  $\alpha = 1$ ) and the nearest centroid classifier (when  $\alpha = 0$ ) with decision rule

$$\mathbb{1} \left\{ \|\hat{\boldsymbol{\mu}}_0 - \mathbf{x}\|_2^2 - \|\hat{\boldsymbol{\mu}}_1 - \mathbf{x}\|_2^2 > 0 \right\} = \mathbb{1} \left\{ \hat{\boldsymbol{\mu}}^T \hat{\mathbf{x}} > 0 \right\}.$$

As suggested by the name, the nearest centroid classifier classifies  $\mathbf{x}$  to the class with nearest sample mean. It is the Bayes classifier for data distributed as (2.3) when  $\Sigma = \mathbf{I}_p$ .

As the previous section has shown,  $\alpha$ -LDA exhibits the greatest improvement in performance among the sampled classifiers, particularly when the data dimensionality  $p$  is on the order of the number of samples  $n$ . This can be attributed to the fact that the LDA weight vector is an explicit function of the sample statistics. Due to estimation noise, there is much to be gained in this regime. We thus pursue an asymptotic study of  $\alpha$ -LDA in growth regime where  $n$  and  $p$  grow at constant rates to each other. Under this growth regime, we derive an asymptotic expression and an estimator for the probability of misclassification of  $\alpha$ -LDA.

### 5.4.1 Asymptotic analysis

In this section we first show that under the following growth regime assumptions

- (a)  $0 < \liminf \frac{p}{n} < \limsup \frac{p}{n} < 1$
- (b)  $\frac{n_i}{n} \rightarrow c_i \in (0, 1), i = 0, 1$
- (c)  $\limsup_p \|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 < \infty$
- (d)  $\limsup_p \|\boldsymbol{\Sigma}_i\|_2 < \infty, i = 0, 1$
- (e)  $\liminf_p \lambda_{\min}(\boldsymbol{\Sigma}_i) > 0, i = 0, 1$

and considering the training set to be random, we are able to derive a DE of the probability of misclassification of the  $\alpha$ -LDA classifier. This may be useful for understanding the behavior of the classifier with synthetic data, for which the statistics are perfectly known. In practice, however, the statistics are unknown. For this reason, we also derive the G-estimator of the probability of misclassification which can be used to tune  $\alpha$ . To proceed with these derivations, we first require an expression for the expected probability of misclassification.

Using (2.4), under the assumption of classes  $\mathcal{C}_0$  and  $\mathcal{C}_1$  are Gaussian with means and covariances  $\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0$  and  $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$  respectively, the probability of mis-

classification of a test point  $\mathbf{x}$  by the  $\alpha$ -LDA classifier is

$$\varepsilon = \pi_0 \Phi \left( \frac{m_0}{\sqrt{\sigma_0^2}} \right) + \pi_1 \Phi \left( -\frac{m_1}{\sqrt{\sigma_1^2}} \right)$$

where  $m_0$ ,  $m_1$ ,  $\sigma_0^2$ , and  $\sigma_1^2$  are the discriminant means and variances conditioned on  $\mathbf{x} \in \mathcal{C}_0$  and  $\mathbf{x} \in \mathcal{C}_1$  respectively. Define  $\rho = \frac{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}}$ . Then

$$m_i = \left( \rho \hat{\boldsymbol{\mu}}^T + \alpha \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \right) \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right), \quad i = 0, 1,$$

and

$$\sigma_i^2 = \left( \rho \hat{\boldsymbol{\mu}}^T + \alpha \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \right) \boldsymbol{\Sigma}_i \left( \rho \hat{\boldsymbol{\mu}}^T + \alpha \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \right)^T, \quad i = 0, 1$$

In the following sections, we present the DEs and G-estimators for both the general case of distinct covariances and the special case of common covariances.

## DE of the probability of misclassification

Formally, the DE of  $\varepsilon$ , denoted by  $\bar{\varepsilon}$ , is a sequence of  $p$  and  $n$  satisfying

$$\varepsilon - \bar{\varepsilon} \xrightarrow{\text{a.s.}} 0$$

under the growth regime assumptions (a)-(e). For sequences  $\bar{m}_0$ ,  $\bar{m}_1$ ,  $\bar{\sigma}_0^2$ , and  $\bar{\sigma}_1^2$  such that

$$\begin{aligned} m_i - \bar{m}_i &\xrightarrow{\text{a.s.}} 0, \quad i = 0, 1 \\ \sigma_i^2 - \bar{\sigma}_i^2 &\xrightarrow{\text{a.s.}} 0, \quad i = 0, 1 \end{aligned} \tag{5.13}$$

under the growth regime assumptions (a)-(e), it is

$$\bar{\varepsilon} = \pi_0 \Phi \left( \frac{\bar{m}_0}{\sqrt{\bar{\sigma}_0^2}} \right) + \pi_1 \Phi \left( -\frac{\bar{m}_1}{\sqrt{\bar{\sigma}_1^2}} \right)$$

(see Lemma 2 in (Niyazi et al., 2020b) for proof). Thus, the DE  $\bar{\varepsilon}$  is itself a function of DEs  $\bar{m}_0$ ,  $\bar{m}_1$ ,  $\bar{\sigma}_0^2$ , and  $\bar{\sigma}_1^2$  which are also functions of only true statistics.

In the following theorem, we state the expressions of  $\bar{m}_0$ ,  $\bar{m}_1$ ,  $\bar{\sigma}_0^2$ , and  $\bar{\sigma}_1^2$  which are used to compute  $\bar{\varepsilon}$ . This is followed by a corollary which corresponds to the special case when  $\Sigma_0 = \Sigma_1 = \Sigma$ .<sup>1</sup> First, define

$$\bar{\mathbf{Q}} := \left( \frac{n_0 - 1}{n - 2} \frac{1}{1 + \tilde{\delta}} \Sigma_0 + \frac{n_1 - 1}{n - 2} \frac{1}{1 + \tilde{\nu}} \Sigma_1 \right)^{-1},$$

$$R_{ij} := \frac{n_{i-1} - 1}{n_{j-1} - 1} [(\mathbf{I}_2 - \Omega)^{-1} \Omega]_{i,j}, \quad i, j = 1, 2,$$

$$[\Omega]_{1j} := \frac{n_{j-1} - 1}{n - 2} \left( \frac{1}{1 + \tilde{\delta}} \right)^2 \frac{1}{n - 2} \text{tr} \{ \Sigma_0 \bar{\mathbf{Q}} \Sigma_{j-1} \bar{\mathbf{Q}} \}, \quad j = 1, 2,$$

$$[\Omega]_{2j} := \frac{n_{j-1} - 1}{n - 2} \left( \frac{1}{1 + \tilde{\nu}} \right)^2 \frac{1}{n - 2} \text{tr} \{ \Sigma_1 \bar{\mathbf{Q}} \Sigma_{j-1} \bar{\mathbf{Q}} \}, \quad j = 1, 2,$$

$$\mathbf{A}_i := \Sigma_i \bar{\mathbf{Q}}, \quad i = 0, 1,$$

$$\tilde{\mathbf{Q}}_i := \bar{\mathbf{Q}} (\mathbf{A}_i + R_{1(i+1)} \mathbf{A}_0 + R_{2(i+1)} \mathbf{A}_1), \quad i = 0, 1,$$

$$\kappa := \frac{\boldsymbol{\mu}^T \bar{\mathbf{Q}} \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \mathbf{A}_0 \} + \frac{1}{n_1} \text{tr} \{ \mathbf{A}_1 \}}{\boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \Sigma_0 \} + \frac{1}{n_1} \text{tr} \{ \Sigma_1 \}},$$

$$\eta := \frac{\left( \frac{1}{1 - \frac{p}{n-2}} \right) \left[ \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right]}{\boldsymbol{\mu}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \Sigma \}},$$

$$\tau := \frac{1}{1 - \frac{p}{n-2}},$$

---

<sup>1</sup>Note in these statements that while technically  $n - 2$  is equivalent to  $n$  asymptotically, we retain the  $n - 2$  in these expressions for increased accuracy in finite dimensions.

and  $\tilde{\delta}$  and  $\tilde{\nu}$  are the results of the fixed point iteration

$$\begin{aligned}\tilde{\delta}^{(k)} &= \frac{1}{n-2} \text{tr} \left\{ \Sigma_0 \left( \frac{n_0-1}{n-2} \frac{1}{1+\tilde{\delta}^{(k-1)}} \Sigma_0 + \frac{n_1-1}{n-2} \frac{1}{1+\tilde{\nu}^{(k-1)}} \Sigma_1 \right)^{-1} \right\} \\ \tilde{\nu}^{(k)} &= \frac{1}{n-2} \text{tr} \left\{ \Sigma_1 \left( \frac{n_0-1}{n-2} \frac{1}{1+\tilde{\delta}^{(k-1)}} \Sigma_0 + \frac{n_1-1}{n-2} \frac{1}{1+\tilde{\nu}^{(k-1)}} \Sigma_1 \right)^{-1} \right\}, \quad k = 1, 2, 3, \dots\end{aligned}$$

for any positive initialization of  $\tilde{\delta}^{(0)}$  and  $\tilde{\nu}^{(0)}$ .

**Theorem 2** (*Distinct covariance DEs*) *The DEs  $\bar{m}_0$ ,  $\bar{m}_1$ ,  $\bar{\sigma}_0^2$ , and  $\bar{\sigma}_1^2$ , satisfying (5.13) under the growth regime assumptions (a)-(e) are given by*

$$\begin{aligned}\bar{m}_i &= (1-\alpha)\kappa \left[ \frac{(-1)^{i+1}}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \Sigma_0 \} - \frac{1}{n_1} \text{tr} \{ \Sigma_1 \} \right) \right] \\ &\quad + \alpha \left[ \frac{(-1)^{i+1}}{2} \boldsymbol{\mu}^T \bar{\mathbf{Q}} \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \mathbf{A}_0 \} - \frac{1}{n_1} \text{tr} \{ \mathbf{A}_1 \} \right) \right], \quad i = 0, 1\end{aligned}$$

and

$$\begin{aligned}\bar{\sigma}_i^2 &= (1-\alpha)^2 \kappa^2 \left[ \boldsymbol{\mu}^T \Sigma_i \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \Sigma_0 \Sigma_i \} + \frac{1}{n_1} \text{tr} \{ \Sigma_1 \Sigma_i \} \right] \\ &\quad + 2\alpha(1-\alpha)\kappa \left[ \boldsymbol{\mu}^T \mathbf{A}_i \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \Sigma_i \mathbf{A}_0 \} + \frac{1}{n_1} \text{tr} \{ \Sigma_i \mathbf{A}_1 \} \right] \\ &\quad + \alpha^2 \left[ \boldsymbol{\mu}^T \tilde{\mathbf{Q}}_i \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \Sigma_0 \tilde{\mathbf{Q}}_i \} + \frac{1}{n_1} \text{tr} \{ \Sigma_1 \tilde{\mathbf{Q}}_i \} \right], \quad i = 0, 1.\end{aligned}$$

**Proof:** See Appendix C.2.1.

**Corollary 2** (*Common covariance DEs*) *The deterministic equivalents  $\bar{m}_0$ ,  $\bar{m}_1$ ,  $\bar{\sigma}_0^2$ , and  $\bar{\sigma}_1^2$  satisfying (5.13) under the growth regime assumptions (a)-(e) are given by*

$$\begin{aligned}\bar{m}_i &= (1-\alpha)\eta \left( \frac{(-1)^{i+1}}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} - \frac{1}{n_1} \right) \text{tr} \{ \Sigma \} \right) \\ &\quad + \alpha \left[ \frac{\tau}{2} \left[ (-1)^{i+1} \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} + \frac{p}{n_0} - \frac{p}{n_1} \right] \right], \quad i = 0, 1\end{aligned}$$

and

$$\begin{aligned} \bar{\sigma}_0^2 = \bar{\sigma}_1^2 = & (1 - \alpha)^2 \eta^2 \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma} \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma}^2 \} \right] \\ & + \alpha^2 \tau^3 \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right] + 2\alpha(1 - \alpha)\tau\eta \left[ \boldsymbol{\mu}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \} \right] \end{aligned}$$

**Proof:** See Appendix C.2.2.

## G-estimator of the probability of misclassification

The G-estimator  $\hat{\varepsilon}$  of the probability of misclassification  $\varepsilon$  is a function of sample statistics  $\hat{\boldsymbol{\mu}}_0$ ,  $\hat{\boldsymbol{\mu}}_1$ ,  $\hat{\boldsymbol{\Sigma}}_0$ , and  $\hat{\boldsymbol{\Sigma}}_1$  such that

$$\hat{\varepsilon} - \varepsilon \xrightarrow{\text{a.s.}} 0$$

under the growth regime assumptions (a)-(f). For sequences  $\hat{m}_0$ ,  $\hat{m}_1$ ,  $\hat{\sigma}_0^2$ , and  $\hat{\sigma}_1^2$ , which are also functions of only sample statistics, such that

$$\begin{aligned} \hat{m}_i - m_i & \xrightarrow{\text{a.s.}} 0, \quad i = 0, 1, \\ \hat{\sigma}_i^2 - \sigma_i^2 & \xrightarrow{\text{a.s.}} 0, \quad i = 0, 1 \end{aligned} \tag{5.14}$$

under the growth regime assumptions (a)-(e), it is

$$\hat{\varepsilon} = \hat{\pi}_0 \Phi \left( \frac{\hat{m}_0}{\sqrt{\hat{\sigma}_0^2}} \right) + \hat{\pi}_1 \Phi \left( -\frac{\hat{m}_1}{\sqrt{\hat{\sigma}_1^2}} \right).$$

The following theorem states the expressions of  $\hat{m}_0$ ,  $\hat{m}_1$ ,  $\hat{\sigma}_0^2$ , and  $\hat{\sigma}_1^2$  which are used to compute  $\hat{\varepsilon}$ . This is followed by a corollary which is specific to the case when  $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$  is assumed. First, define

$$\lambda_i = \frac{\frac{1}{n-2} \text{tr} \{ \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{\Sigma}}_i^{-1} \}}{1 - \frac{1}{n-2} \text{tr} \{ \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{\Sigma}}_i^{-1} \}}, \quad i = 0, 1.$$

**Theorem 3** (*Distinct covariance G-estimators*) The G-estimators  $\hat{m}_0$ ,  $\hat{m}_1$ ,  $\hat{\sigma}_0^2$ , and  $\hat{\sigma}_1^2$ , satisfying (5.14) under the growth regime assumptions (a)-(e) are given by

$$\hat{m}_i = (-1)^{i+1} \left[ (1 - \alpha) \rho \left( \frac{1}{2} \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}} - \frac{1}{n_i} \text{tr} \{ \hat{\boldsymbol{\Sigma}}_i \} \right) + \alpha \left( \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} - \frac{n-2}{n_i} \lambda_i \right) \right], \quad i = 0, 1$$

and

$$\begin{aligned} \hat{\sigma}_i^2 &= (1 - \alpha)^2 \rho^2 \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{\mu}} + 2\alpha(1 - \alpha) \rho (1 + \lambda_i) \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \\ &\quad + \alpha^2 (1 + \lambda_i)^2 \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Sigma}}_i \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}, \quad i = 0, 1 \end{aligned}$$

**Proof:** See Appendix C.3.1.

**Corollary 3** (*Common covariance G-estimators*) The G-estimators  $\hat{m}_0$ ,  $\hat{m}_1$ ,  $\hat{\sigma}_0^2$ , and  $\hat{\sigma}_1^2$ , satisfying (5.14) under the growth regime assumptions (a)-(e) are given by

$$\begin{aligned} \hat{m}_i &= \frac{(-1)^{i+1}}{2} \left( \rho \hat{\boldsymbol{\mu}}^T + \alpha \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \right) \hat{\boldsymbol{\mu}} \\ &\quad + (-1)^{i+1} \left[ \rho(\alpha - 1) \frac{1}{n_i} \text{tr} \{ \hat{\boldsymbol{\Sigma}} \} - \alpha \frac{\frac{p}{n_i}}{1 - \frac{p}{n-2}} \right], \quad i = 0, 1 \end{aligned}$$

and

$$\hat{\sigma}_0^2 = \hat{\sigma}_1^2 = \rho^2 (1 - \alpha)^2 \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\mu}} + \alpha^2 \tau^2 \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} + 2\alpha \rho (1 - \alpha) \tau \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}$$

**Proof:** See Appendix C.3.2.

Notice that  $\hat{\varepsilon}$  is a function of the sample statistics. It estimates the probability of misclassification without the need for additional testing data and it is much more computationally efficient than the cross-validation procedure. In the next section, we show how to use  $\hat{\varepsilon}$  for the purpose of tuning the  $\alpha$  parameter.

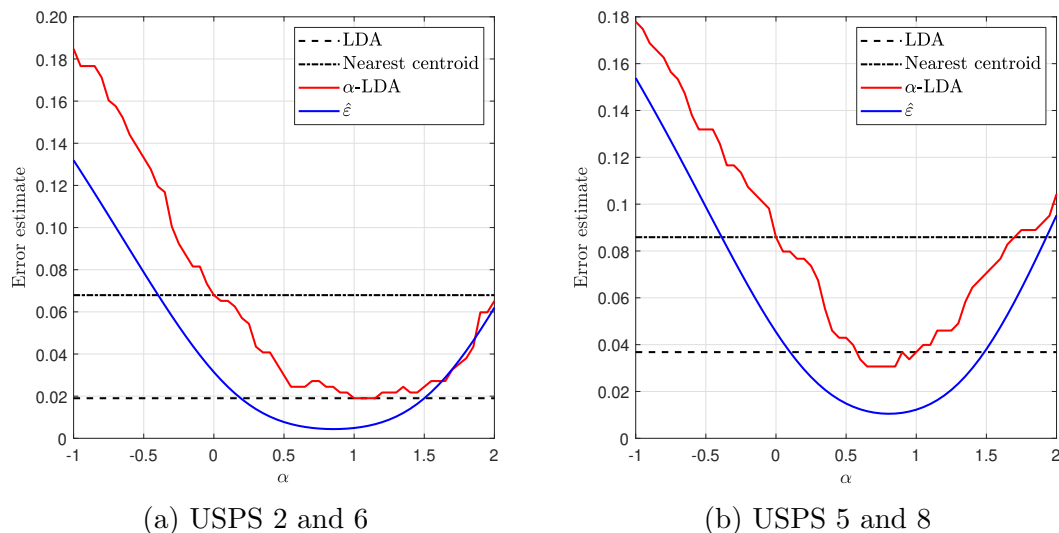


Figure 5.9: Plots of testing error estimates of classifying USPS digit pairs for LDA, the nearest centroid, and  $\alpha$ -LDA as well as the G-estimator  $\hat{\varepsilon}$  of the  $\alpha$ -LDA expected testing error.

### 5.4.2 Tuning the $\alpha$ -LDA parameter

In this section,  $\alpha$ -LDA is applied to real data. The objective is to show how  $\alpha$ -LDA performs as compared to LDA and the nearest centroid classifier on real data, as well as to demonstrate the use of the G-estimator  $\hat{\varepsilon}$  in tuning the  $\alpha$  parameter. We consider binary classification of digit pairs from the USPS dataset (Le Cun et al., 1990) and phoneme pairs from the dataset (Hastie et al., 1995). For each problem, we train and test LDA, nearest centroid, and  $\alpha$ -LDA on the relevant dataset. The empirical errors are plotted against varying  $\alpha$ . Also plotted is the G-estimator  $\hat{\varepsilon}$  of the error of  $\alpha$ -LDA.<sup>2</sup>

Figure 5.9 shows the results on two digit pairs from the USPS dataset. As mentioned in Section 5.3.2, this dataset consists of grayscale images of handwritten digits 0 – 9 encoded as 256-dimensional vectors.

For Figure 5.9a, we use the digit pair ‘2’ and ‘6’. Overall, there are  $n = 1395$  total training vectors and 368 total testing vectors corresponding to this digit pair. The figure shows that LDA achieves the lowest empirical error on this digit pair. This performance is matched by  $\alpha$ -LDA at  $\alpha = 1$ . Although  $\hat{\varepsilon}$  does

<sup>2</sup>Note that for these particular datasets, the two G-estimators almost match. Out of the two, the G-estimator which assumes common covariances is plotted.

not exactly match the empirical error, for parameter tuning it suffices that it follows the same trend. In this case, if we had directly used  $\hat{\varepsilon}$  to tune the  $\alpha$  parameter, we would have set it to  $\alpha = 0.85$ . This setting results in an increase of merely 0.0054 in error compared to the optimal setting. For more sensitive applications, the parameter setting suggested by the G-estimator may be used as a starting point from which to search for the optimal  $\alpha$  using a more accurate (but computationally-intensive) method.

For Figure 5.9b, we use the digit pair ‘5’ and ‘8’. Overall, there are  $n = 1098$  total training vectors and 326 total testing vectors corresponding to this digit pair. In this case,  $\alpha$ -LDA achieves the lowest error of 0.0307 at  $\alpha = 0.65$ . This is a 16.6% decrease in error relative to LDA which has an error rate of 0.0368. If we had directly used  $\hat{\varepsilon}$  to tune the  $\alpha$  parameter, we would have set it to  $\alpha = 0.8$ . This setting incurs no loss in accuracy. Notice this dataset has less training samples than the last one. The increased estimation noise explains why  $\alpha$ -LDA is able to provide a performance advantage over LDA.

Figure 5.10 considers a phoneme pair. The phoneme dataset consists of a total of 4509 instances of digitized speech vectors of the five phonemes ‘aa’, ‘ao’, ‘dcl’, ‘iy’, and ‘sh’, having 256 features each. All 1717 instances of the phonemes ‘ao’ and ‘aa’ (which are the closest in pronunciation) were extracted in order to construct this binary classification problem. As the dataset is not pre-divided into training and testing sets, the splitting was performed randomly. We take advantage of this to construct a classification problem in which  $n$  is not much greater than  $p$ . A training set consisting of 400 samples is randomly extracted from the full set of ‘aa’ and ‘ao’ phonemes according to the same proportions. This leaves 1317 samples for testing. Based on the simulations from the previous section, we expect to observe a much greater performance gain in this scenario compared to Figure 5.9.

Figure 5.10 shows that, as expected,  $\alpha$ -LDA significantly outperforms LDA with an error of 0.224 corresponding to the former compared to 0.3083 corre-

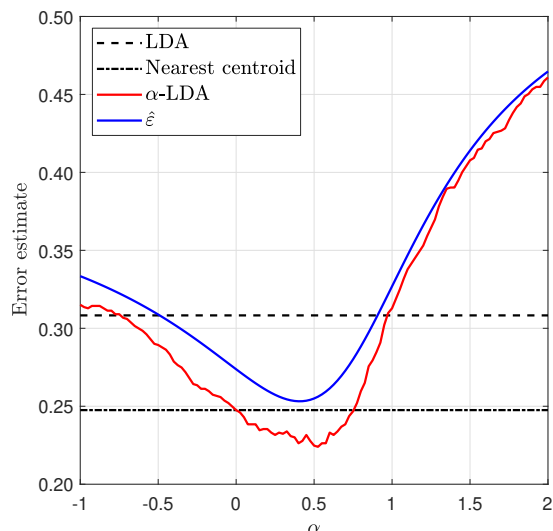


Figure 5.10: Plots of testing error estimates of classifying phonemes ‘aa’ and ‘ao’ for LDA, the nearest centroid, and  $\alpha$ -LDA as well as the G-estimator  $\hat{\epsilon}$  of the  $\alpha$ -LDA expected testing error.

sponding to the latter. It achieves a 27.3% decrease in error at  $\alpha = 0.525$ . In this case, it seems that the data leans more towards an isotropic covariance structure, as nearest centroid performs better than LDA. Even so,  $\alpha = 0$  is not optimal. Thus,  $\alpha$ -LDA provides the best balance between both of these classifiers. Lastly, the G-estimator points towards an  $\alpha$  setting of 0.4. Using this setting incurs an increase in error of just 0.0023 relative to the optimal setting.

In the next section, we show how the proposed  $\alpha$  parameterization of the weight vector of a binary linear classifier can be extended to neural networks in the context of transfer learning.

## 5.5 Transfer learning application

As contemporary deep neural networks typically have millions of parameters, they need to be trained on massive datasets. One example of such a dataset is the ImageNet database which, as of the time of writing, consists of a total of 14,197,122 images belonging to 1000 different object classes. Given a new classification task, one may not have enough data to train a deep neural network from scratch. In this case, transfer learning - which takes advantage of networks pretrained on related data - might be useful.

The idea behind transfer learning is to leverage pretrained neural networks for a different task than the one for which they have been trained and for which the sample size is relatively small. This is done by preserving the pretrained weights in the earlier layers and retraining only the later layers on the smaller dataset. For example, a GoogLeNet architecture pretrained on ImageNet can be adapted to operate as a binary classifier which distinguishes between two specific species of flowers *not* in ImageNet. If the flower dataset is very small, only the last layer having trainable weights, i.e., the last *learnable* layer, and the final softmax layer are replaced with the appropriate configurations for binary classification. During training, all layers are frozen except for the last learnable layer so that only this layer's weights are updated. With more data, even more of the learnable layers could be unfrozen and retrained as well. The intuition behind all of this is that earlier layers of the deep neural network learn basic features while the later layers fine-tune these features, and so only the later layers need to be retrained for the specific dataset at hand (Ng, n.d.). By taking advantage of pretraining, transfer learning speeds up training time and saves energy and valuable computational resources (Chahal and Toner, 2021).

Sometimes the new dataset is too small even for transfer learning. Reference (Inc., 2023) recommends that for a dataset of less than about 20 images per class, it may be better to apply the pretrained neural network to the data as a feature extractor to be used in conjunction with a simple classifier such as the SVM. In this section, we propose to go ahead with the transfer learning even for very small sample sizes, with the expectation that the retrained weights in the later layers be non-optimal. We then apply weight vector tuning to these weights. In the following section, we describe the simulation setup in more detail.

### 5.5.1 Simulation setup

As explained in the previous section, when only a small dataset is available for retraining the later layers in transfer learning, the obtained weights may not be

optimal. In this chapter, we proposed a method for weight vector modification. More specifically, given two-class data with sample means  $\hat{\boldsymbol{\mu}}_0$  and  $\hat{\boldsymbol{\mu}}_1$  belonging to class  $\mathcal{C}_0$  and  $\mathcal{C}_1$  respectively, and given an existing weight vector  $\mathbf{w}$ , the weight vector is tuned through a scalar parameter  $\alpha$  to obtain the modified weight vector  $\mathbf{w}'$

$$\mathbf{w}' = \frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}} + \alpha \mathbf{P}_{\hat{\boldsymbol{\mu}}} \mathbf{w}, \quad (5.15)$$

where  $\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$  and  $\mathbf{P}_{\hat{\boldsymbol{\mu}}} = \left( \mathbf{I} - \frac{\hat{\boldsymbol{\mu}} \hat{\boldsymbol{\mu}}^T}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \right)$ . Note that  $\alpha = 1$  recovers the original weight vector  $\mathbf{w}$ .

In this section, we apply the weight vector tuning technique to a pretrained deep neural network which has been adapted for binary classification and trained using the transfer learning technique in order to check for any improvements in the classification performance as compared to  $\alpha = 1$ . We will take networks pretrained on the ImageNet dataset and apply transfer learning to them to use them for a binary classification problem constructed using built-in MATLAB image datasets for deep learning. Only the last learnable layer is retrained. Afterwards, the newly obtained weights from the last learnable layer are tuned based on (5.15) with different  $\alpha$ . The error rate on the test set is then plotted against  $\alpha$ .

Note that depending on the neural network architecture, the last learnable layer may be a fully-connected layer or it may be a convolutional layer. The fully-connected layer has the form  $\mathbf{w}^T \mathbf{x} + b$ , whereas the convolutional layer does not. Therefore, weight vector tuning can be applied only to neural networks which have a fully-connected layer as their last learnable layer. As an example, GoogLeNet and ResNet-50 are two deep neural network architectures whose last learnable layers are fully-connected layers, whereas SqueezeNet is an architecture whose last learnable layer is a convolutional layer. Additionally, to apply transfer learning to a pretrained neural network as described, new layers with a modified configuration need to be created in place of the last learnable layer and the softmax layer (which gives probabilities for each class). In our simulations, these configurations must allow for binary classification, so we change the number of

nodes to two in the new fully connected layer. The number of classes is learned automatically by MATLAB in the softmax layer during training. Finally, it is important to note that the means,  $\hat{\boldsymbol{\mu}}_0$  and  $\hat{\boldsymbol{\mu}}_1$ , computed for (5.15) are those of the features input to the last fully-connected layer for each class NOT of the raw images input to network.

## 5.5.2 Results

In this section we present the results of the simulation on several deep neural network architectures pretrained on the ImageNet dataset and transfer-learned on two-class data extracted from several of the built-in MATLAB image datasets for deep learning available at <https://www.mathworks.com/help/deeplearning/ug/data-sets-for-deep-learning.html>. The data is divided into training and testing sets in such a way as to artificially construct a small sample problem. The resulting binary classification problems are as follows:

- ‘daisy’ and ‘sunflower’ from the ‘Flowers’ dataset with a total of 133 training samples and 1199 test samples;
- ‘rose’ and ‘tulip’ from the ‘Flowers’ dataset with a total of 144 training samples and 1296 test samples;
- and ‘pizza’ and ‘hamburger’ from the ‘Example Food Images’ dataset with a total of 27 training samples and 510 test samples.

Note that ‘sunflower’, ‘rose’, ‘tulip’, and ‘hamburger’ are not amongst the ImageNet classes and therefore none of these classification problems are feasible with the unmodified pretrained network. The pretrained deep neural network architectures considered are GoogLeNet and ResNet-50.

Figures 5.11 to 5.13 plot the testing error rates of the weight vector tuned transfer learned GoogLeNet architecture for the ‘daisy’/‘sunflower’, ‘rose’/‘tulip’, and ‘pizza’/‘hamburger’ binary classification problems, respectively. In Figure 5.11, a very slight reduction is observed for the ‘daisy’/‘sunflower’ dataset of

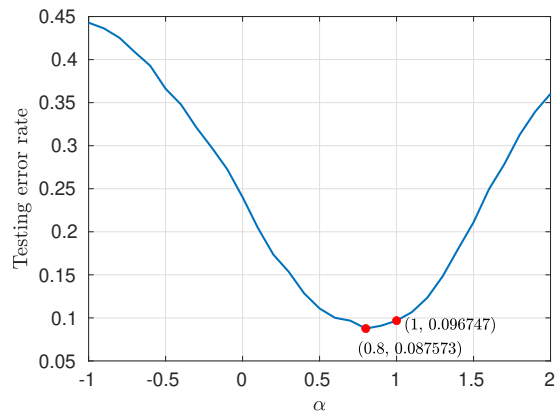


Figure 5.11: Plot of testing error rate of the weight vector tuned GoogLeNet which was transfer learned on ‘daisy’ and ‘sunflower’ images.

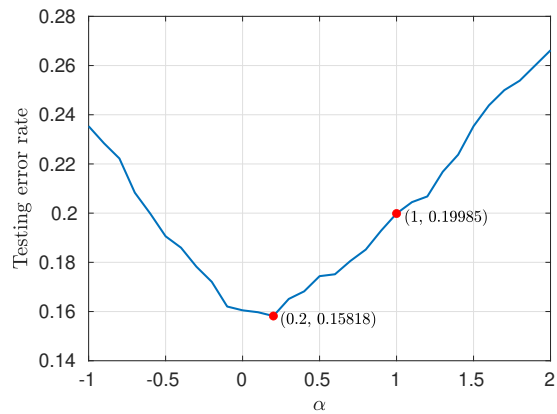


Figure 5.12: Plot of testing error rate of the weight vector tuned GoogLeNet which was transfer learned on ‘rose’ and ‘tulip’ images.

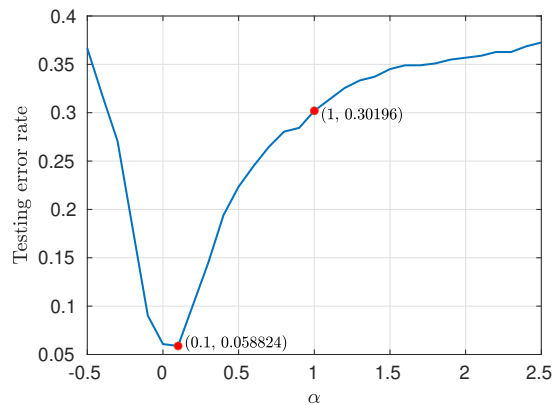


Figure 5.13: Plot of testing error rate of the weight vector tuned GoogLeNet which was transfer learned on ‘pizza’ and ‘hamburger’ images.

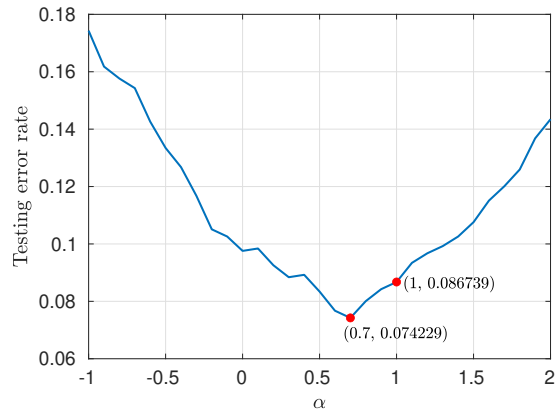


Figure 5.14: Plot of testing error rate of the weight vector tuned ResNet-50 which was transfer learned on ‘daisy’ and ‘sunflower’ images.

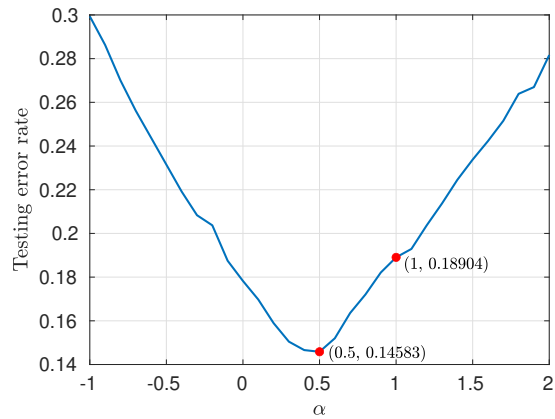


Figure 5.15: Plot of testing error rate of the weight vector tuned ResNet-50 which was transfer learned on ‘rose’ and ‘tulip’ images.

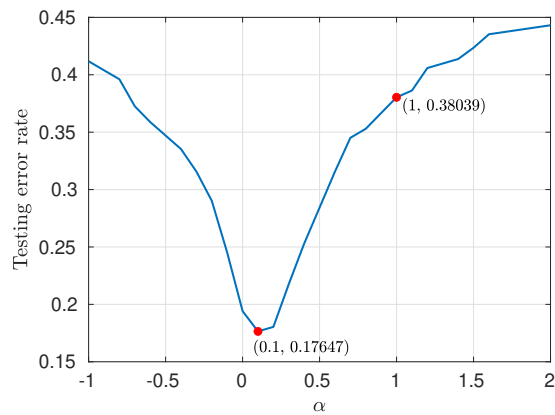


Figure 5.16: Plot of testing error rate of the weight vector tuned ResNet-50 which was transfer learned on ‘pizza’ and ‘hamburger’ images.

about 0.01 reduction in error rate at  $\alpha = 0.8$ . A moderate reduction in error rate of around 0.04 at  $\alpha = 0.2$  is observed for the ‘rose’/‘tulip’ dataset in Figure 5.12. A huge reduction in error rate of around 0.25 at  $\alpha = 0.1$  is observed for the ‘pizza’/‘hamburger’ dataset in Figure 5.13. Similar trends are observed in Figures 5.14 to 5.16 which plot the testing error rates of the weight vector tuned transfer learned ResNet-50 architecture for the ‘daisy’/‘sunflower’, ‘rose’/‘tulip’, and ‘pizza’/‘hamburger’ binary classification problems, respectively. In particular, Figure 5.16 shows that weight vector tuning is able to reduce the error rate by about 0.21 at  $\alpha = 0.1$  for the ‘pizza’/‘hamburger’ dataset. This showcases the potential for the weight vector tuning technique to improve performance in the context of deep neural networks beyond simple classifiers such as LDA and linear SVM.

## Chapter 6

### Concluding Remarks

This dissertation considered the performance analysis and design of linear classifiers for high-dimensional, small sample data with a focus on LDA-based methods. More specifically, we went through a journey from the study of random projection based variants of the LDA classifier in Chapter 3 to designing a rotationally-invariant precision estimator for use specifically with LDA in Chapter 4 and finally to the development of a weight vector tuning technique for a generic binary linear classifier in Chapter 5. The common theme in all of these works is our leveraging of asymptotic RMT tools for the derivation of the limits of quantities of interest as well as consistent estimators of those quantities in the asymptotic regime where both data dimension and sample size grow commensurately. As a result of this analysis we are able to extract insights into the behavior of the classifiers we study as well as develop more computationally efficient methods for tuning their hyperparameters when faced with data of a high-dimensional, small sample nature.

Possible future work is an extension of the proposed weighted shrinkage scheme of Chapter 4 to sample spectra consisting of multiple bulks or even multiple bulks plus a finite number of spiked eigenvalues. In such cases, shrinking each bulk (and spike) individually using its own set of optimized weights should yield improved performance.

## REFERENCES

- U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.
- E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn *et al.*, "Use of proteomic patterns in serum to identify ovarian cancer," *The lancet*, vol. 359, no. 9306, pp. 572–577, 2002.
- D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer cell*, vol. 1, no. 2, pp. 203–209, 2002.
- T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, vol. 286, no. 5439, pp. 531–537, 1999.
- S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, F. Mandelli, J. Ritz, and R. Foa, "Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," *Blood*, vol. 103, no. 7, pp. 2771–2778, 2004.
- F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 1994 IEEE workshop on applications of computer vision*. IEEE, 1994, pp. 138–142.
- P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- A. Martinez and R. Benavente, "The ar face database: Cvc technical report, 24," 1998.
- D. B. Graham and N. M. Allinson, "Characterising virtual eigensignatures for general purpose face recognition," in *Face recognition: from theory to applications*. Springer, 1998, pp. 446–456.
- M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

- T. Mehmood and M. Iqbal, “Ftir fingerprints discriminate ionic liquids’ antibacterial activity,” *Chemometrics and Intelligent Laboratory Systems*, vol. 208, p. 104200, 2021.
- A. Kanwal, T. Mehmood, and M. M. Butt, “Pls and kernel svm based hybrid classifier for discriminating ftir spectrum data with limited sample size,” *Chemometrics and Intelligent Laboratory Systems*, vol. 215, p. 104365, 2021.
- Z. Liao, “A random matrix framework for large dimensional machine learning and neural networks,” Ph.D. dissertation, Université Paris-Saclay, 2019.
- O. Ledoit and M. Wolf, “The power of (non-) linear shrinking: A review and guide to covariance matrix estimation,” *Journal of Financial Econometrics*, vol. 20, no. 1, pp. 187–218, 2022.
- D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, “Microarray data analysis: from disarray to consolidation and consensus,” *Nature reviews genetics*, vol. 7, no. 1, pp. 55–65, 2006.
- X. Tong, L. Xia, J. Wang, and Y. Feng, “Neyman-pearson classification: parametrics and sample size requirement,” *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 380–427, 2020.
- H. Sifaou, A. Kammoun, and M.-S. Alouini, “High-dimensional linear discriminant analysis classifier for spiked covariance model,” *Journal of Machine Learning Research*, vol. 21, pp. 1–24, 2020.
- S. Huang, T. Tong, and H. Zhao, “Bias-corrected diagonal discriminant rules for high-dimensional classification,” *Biometrics*, vol. 66, no. 4, pp. 1096–1106, 2010.
- H. Li, W. Luo, Z. Bai, H. Zhou, and Z. Pu, “Spectrally-corrected and regularized linear discriminant analysis for spiked covariance model,” *arXiv preprint arXiv:2210.03859*, 2022.
- R. J. Durrant and A. Kabán, “Random projections as regularizers: learning a linear discriminant from fewer observations than dimensions,” *Machine Learning*, vol. 99, no. 2, pp. 257–286, 2015.
- T. I. Cannings and R. J. Samworth, “Random-projection ensemble classification,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 79, no. 4, pp. 959–1035, 2017.
- Q. Lu and X. Qiao, “Sparse fisher’s linear discriminant analysis for partially labeled data,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 11, no. 1, pp. 17–31, 2018.
- D. M. Witten and R. Tibshirani, “Penalized classification using fisher’s linear discriminant,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 73, no. 5, pp. 753–772, 2011.
- A. Chatterjee, S. Mazumder, and K. Das, “Functional classwise principal com-

- ponent analysis: a classification framework for functional data analysis,” *Data Mining and Knowledge Discovery*, vol. 37, no. 2, pp. 552–594, 2023.
- R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, “Diagnosis of multiple cancer types by shrunken centroids of gene expression,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 10, pp. 6567–6572, 2002.
- P. Xu, G. N. Brock, and R. S. Parrish, “Modified linear discriminant analysis approaches for classification of high-dimensional microarray data,” *Computational Statistics & Data Analysis*, vol. 53, no. 5, pp. 1674–1687, 2009.
- P.-H. Huynh, V. H. Nguyen, and T.-N. Do, “Random ensemble oblique decision stumps for classifying gene expression data,” in *Proceedings of the 9th International Symposium on Information and Communication Technology*, 2018, pp. 137–144.
- T. Mehmood, A. Kanwal, and M. M. Butt, “Naive bayes combined with partial least squares for classification of high dimensional microarray data,” *Chemo-metrics and Intelligent Laboratory Systems*, vol. 222, p. 104492, 2022.
- H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.
- P.-H. Huynh, V.-H. Nguyen, and T.-N. Do, “A combined enhancing and feature extraction algorithm to improve learning accuracy for gene expression classification,” in *Future Data and Security Engineering: 6th International Conference, FDSE 2019, Nha Trang City, Vietnam, November 27–29, 2019, Proceedings 6*. Springer, 2019, pp. 255–273.
- J. Hua, W. D. Tembe, and E. R. Dougherty, “Performance of feature-selection methods in the classification of high-dimension data,” *Pattern Recognition*, vol. 42, no. 3, pp. 409–424, 2009.
- A. Sharma and K. K. Paliwal, “Linear discriminant analysis for the small sample size problem: An overview,” *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 3, pp. 443–454, Jun 2015. [Online]. Available: <https://doi.org/10.1007/s13042-013-0226-9>
- H. Nakouri, “Two-dimensional subclass discriminant analysis for face recognition,” *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 109–117, 2021.
- C. E. Thomaz, E. C. Kitani, and D. F. Gillies, “A maximum uncertainty lda-based approach for limited sample size problems—with application to face recognition,” *Journal of the Brazilian Computer Society*, vol. 12, pp. 7–18, 2006.
- F. Song, J. Yang, and S. Liu, “Large margin linear projection and face recognition,” *Pattern recognition*, vol. 37, no. 9, pp. 1953–1955, 2004.
- F. Song, D. Zhang, Q. Chen, and J. Wang, “Face recognition based on a novel linear discriminant criterion,” *Pattern analysis and applications*, vol. 10, pp.

- 165–174, 2007.
- M. Murtaza, M. Sharif, M. Raza, and J. Shah, “Face recognition using adaptive margin fisher’s criterion and linear discriminant analysis,” *International Arab Journal of Information Technology*, vol. 11, no. 2, pp. 1–11, 2014.
- A. Dey, S. Chowdhury, and J. K. Sing, “Feature extraction using fuzzy generalized two-dimensional inverse lda with gaussian probabilistic distribution and face recognition,” in *Advanced Computational and Communication Paradigms: Proceedings of International Conference on ICACCP 2017, Volume 2*. Springer, 2018, pp. 553–561.
- J. Xu, Q. Xu, L. Yi, C.-O. Chan, and D. K.-W. Mok, “Correlation-assisted nearest shrunken centroid classifier with applications for high dimensional spectral data,” *Journal of Chemometrics*, vol. 30, no. 1, pp. 37–45, 2016.
- T.-S. Lim, W.-Y. Loh, and Y.-S. Shih, “A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms,” *Machine learning*, vol. 40, no. 3, pp. 203–228, 2000.
- D. J. Hand, “Classifier technology and the illusion of progress,” *Statistical science*, vol. 21, no. 1, pp. 1–14, 2006.
- A. Müller and M. Debbah, “Random matrix theory tutorial-Introduction to deterministic equivalents,” *Traitement du signal*, vol. 33, no. 2-3, pp. 223–248, 2016.
- J. Najim and R. Couillet, “Random matrix approach for machine learning,” Presentation for the Winter Enrichment Program at the King Abdullah University of Science and Technology, Jan 2018.
- M. Krishnapur, “Random matrix theory,” Lecture notes from the Indian Institute of Science, 2011.
- G. W. Anderson, A. Guionnet, and O. Zeitouni, “An introduction to random matrices,” 2010.
- R. Couillet and M. Debbah, *Random matrix methods for wireless communications*. Cambridge University Press, 2011.
- A. M. Tulino, S. Verdú *et al.*, “Random matrix theory and wireless communications,” *Foundations and Trends® in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.
- T. L. Marzetta, G. H. Tucci, and S. H. Simon, “A random matrix-theoretic approach to handling singular covariance estimates,” *IEEE Transactions on Information Theory*, vol. 57, no. 9, pp. 6256–6271, 2011.
- L. B. Niyazi, A. Kammoun, H. Dahrouj, M.-S. Alouini, and T. Y. Al-Naffouri, “Asymptotic analysis of an ensemble of randomly projected linear discriminants,” *arXiv preprint arXiv:2004.08217*, 2020.
- R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals*

- of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.
- J.-P. Vert, “Link between lda and ols,” Jun 2011. [Online]. Available: <https://members.cbio.mines-paristech.fr/~jvert/svn/tutorials/course/1102Senegal/solution.pdf>
- J. Hoydis, “Random matrix theory for advanced communication systems.” Ph.D. dissertation, Supélec, 2012.
- W. Hachem, P. Loubaton, J. Najim, and P. Vallet, “On bilinear forms based on the resolvent of large random matrices,” in *Annales de l’IHP Probabilités et statistiques*, vol. 49, no. 1, 2013, pp. 36–63.
- F. Benaych-Georges and R. Couillet, “Spectral analysis of the gram matrix of mixture models,” *ESAIM: Probability and Statistics*, vol. 20, pp. 217–237, 2016.
- A. Kammoun, L. Sanguinetti, M. Debbah, and M.-S. Alouini, “Asymptotic analysis of RZF in large-scale MU-MIMO systems over rician channels,” *IEEE Transactions on Information Theory*, vol. 65, no. 11, pp. 7268–7286, 2019.
- E. Bingham and H. Mannila, “Random projection in dimensionality reduction: Applications to image and text data,” in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’01. New York, NY, USA: ACM, 2001, pp. 245–250. [Online]. Available: <http://doi.acm.org/10.1145/502512.502546>
- R. J. Durrant and A. Kabán, “Random projections as regularizers: Learning a linear discriminant ensemble from fewer observations than dimensions,” in *Proceedings of the Asian Conference on Machine Learning*, vol. 29. JMLR, 2013, pp. 17–32. [Online]. Available: <http://jmlr.org/proceedings/papers/v29/Durrant13.html>GoogleScholar
- J. Matoušek, “Lecture notes on metric embeddings,” Technical report, ETH Zürich, Tech. Rep., 2013.
- A. Schclar and L. Rokach, “Random projection ensemble classifiers,” in *International Conference on Enterprise Information Systems*. Springer, 2009, pp. 309–316.
- D. Peressutti, W. Bai, T. Jackson, M. Sohal, A. Rinaldi, D. Rueckert, and A. King, “Prospective identification of CRT super responders using a motion atlas and random projection ensemble learning,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 493–500.
- L. B. Niyazi, A. Kammoun, H. Dahrouj, M.-S. Alouini, and T. Y. Al-Naffouri, “Asymptotic analysis of an ensemble of randomly projected linear discriminants,” *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 3, pp. 914–930, 2020.
- A. Kabán, “On compressive ensemble induced regularisation: How close is the

- finite ensemble precision matrix to the infinite ensemble?” in *International Conference on Algorithmic Learning Theory*. PMLR, 2017, pp. 617–628.
- , “Sufficient ensemble size for random matrix theory-based handling of singular covariance matrices,” *Analysis and Applications*, vol. 18, no. 05, pp. 929–950, 2020.
- T. I. Cannings, “Random projections: Data perturbation for classification problems,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 13, no. 1, p. e1499, 2021.
- G. C. Cawley and N. L. Talbot, “On over-fitting in model selection and subsequent selection bias in performance evaluation,” *The Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, 2010.
- J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.
- P. Mesejo, D. Pizarro, A. Abergel, O. Rouquette, S. Beorchia, L. Poincloux, and A. Bartoli, “Computer-aided classification of gastrointestinal lesions in regular colonoscopy,” *IEEE transactions on medical imaging*, vol. 35, no. 9, pp. 2051–2063, 2016.
- T. Hastie, A. Buja, and R. Tibshirani, “Penalized discriminant analysis,” *The Annals of Statistics*, pp. 73–102, 1995.
- J. H. Friedman, “Regularized discriminant analysis,” *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.
- C. Stein, “Lectures on the theory of estimation of many parameters,” *Journal of Soviet Mathematics*, vol. 34, pp. 1373–1403, 1986.
- O. Ledoit and M. Wolf, “Nonlinear shrinkage estimation of large-dimensional covariance matrices,” *The Annals of Statistics*, vol. 40, no. 2, pp. 1024–1060, 2012.
- Y. Guo, T. Hastie, and R. Tibshirani, “Regularized linear discriminant analysis and its application in microarrays,” *Biostatistics*, vol. 8, no. 1, pp. 86–100, 2007.
- L. Yang, M. R. McKay, and R. Couillet, “High-dimensional mvdr beamforming: Optimized solutions based on spiked random matrix models,” *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1933–1947, 2018.
- O. Ledoit and M. Wolf, “Quadratic shrinkage for large covariance matrices,” *Bernoulli*, vol. 28, no. 3, pp. 1519–1547, 2022.
- , “Analytical nonlinear shrinkage of large-dimensional covariance matrices,” *The Annals of Statistics*, vol. 48, no. 5, pp. 3043–3065, 2020.
- A. F. Marquand and S. M. Kia, “Chapter 5 - linear methods for classification,” in *Machine Learning*, A. Mechelli and S. Vieira, Eds. Academic Press, 2020, pp. 83 – 100. [Online]. Available: <http://>

- [//www.sciencedirect.com/science/article/pii/B9780128157398000055](http://www.sciencedirect.com/science/article/pii/B9780128157398000055)
- J. Wen, J. Zhu, T. Xue, J. Cang, L. Wei, Q. Nie, M. Zeng, Z. Zeng, H. Ma, J. Li *et al.*, “Performance of linear classification algorithms on  $\alpha/\gamma$  discrimination for labr3: Ce scintillation detectors with various pulse digitizer properties,” *Journal of Instrumentation*, vol. 15, no. 02, p. P02004, 2020.
- D. P. Berry, T. Pabiou, R. Fanning, R. D. Evans, and M. M. Judge, “Linear classification scores in beef cattle as predictors of genetic merit for individual carcass primal cut yields,” *Journal of animal science*, vol. 97, no. 6, pp. 2329–2341, 2019.
- G. S. Randhawa, M. P. Soltysiak, H. El Roz, C. P. de Souza, K. A. Hill, and L. Kari, “Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study,” *Plos one*, vol. 15, no. 4, p. e0232391, 2020.
- P. Ashok, T. Brázdil, K. Chatterjee, J. Křetínský, C. H. Lampert, and V. Toman, “Strategy representation by decision trees with linear classifiers,” in *International Conference on Quantitative Evaluation of Systems*. Springer, 2019, pp. 109–128.
- R. Alanni, J. Hou, H. Azzawi, and Y. Xiang, “A novel gene selection algorithm for cancer classification using microarray datasets,” *BMC medical genomics*, vol. 12, no. 1, p. 10, 2019.
- G.-X. Yuan, C.-H. Ho, and C.-J. Lin, “Recent advances of large-scale linear classification,” *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2584–2603, 2012.
- R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern classification*. Wiley, 2001.
- C. Wang, B. Jiang *et al.*, “On the dimension effect of regularized linear discriminant analysis,” *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 2709–2742, 2018.
- A. Zollanvari, M. Abdirash, A. Dadlani, and B. Abibullaev, “Asymptotically bias-corrected regularized linear discriminant analysis for cost-sensitive binary classification,” *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1300–1304, 2019.
- Q. Mai, H. Zou, and M. Yuan, “A direct approach to sparse discriminant analysis in ultra-high dimensions,” *Biometrika*, vol. 99, no. 1, pp. 29–42, 2012.
- E. W. Weisstein, “Hypersphere Point Picking,” *From MathWorld—A Wolfram Web Resource*, 2017, available at <http://mathworld.wolfram.com/HyperspherePointPicking.html>.
- L. Bottou and C.-J. Lin, “Support vector machine solvers,” *Large scale kernel machines*, vol. 3, no. 1, pp. 301–320, 2007.
- Y. Le Cun, O. Matan, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. Jacket, and H. S. Baird, “Handwritten zip code recognition

- with multilayer networks,” in *[1990] Proceedings. 10th International Conference on Pattern Recognition*, vol. 2. IEEE, 1990, pp. 35–40.
- A. Ng, “Transfer learning,” Coursera, n.d., accessed 8/31/2023. [Online]. Available: <https://www.coursera.org/learn/convolutional-neural-networks/lecture/4THzO/transfer-learning>
- H. Chahal and H. Toner, “‘Small data’ are also crucial for machine learning,” *Scientific American*, 2021.
- T. M. Inc., Natick, Massachusetts, United States, 2023. [Online]. Available: <https://www.mathworks.com/help/deeplearning/ug/pretrained-convolutional-neural-networks.html>
- W. Hachem, O. Khorunzhiy, P. Loubaton, J. Najim, and L. Pastur, “A new approach for mutual information analysis of large dimensional multi-antenna channels,” *IEEE Transactions on Information Theory*, vol. 54, no. 9, pp. 3987–4004, 2008.
- A. W. Van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 2000, vol. 3.
- G. Witt, “Moody’s correlated binomial default distribution,” *Moody’s Investor Service, Special Report, August*, 2004.
- P. Borjesson and C.-E. Sundberg, “Simple approximations of the error function  $q(x)$  for communications applications,” *IEEE Transactions on Communications*, vol. 27, no. 3, pp. 639–643, 1979.
- F. Rubio and X. Mestre, “Consistent reduced-rank LMMSE estimation with a limited number of samples per observation dimension,” *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 2889–2902, 2009.
- D. Fischer, “Uniformly bounded derivative implies uniform convergence,” Mathematics Stack Exchange, July 2014. [Online]. Available: <https://math.stackexchange.com/q/875205>
- A. Kammoun and M.-S. Alouini, “On the smallest eigenvalue of general correlated gaussian matrices,” *arXiv preprint arXiv:1412.8340*, 2014.

## APPENDICES

### Appendix A

#### Proofs for Chapter 3

##### A.1 Single RP-LDA class-conditional discriminant statistics

###### A.1.1 Means

In this section, we derive the DE of the quantity  $m_i(1)$ ,  $i = 0, 1$ , defined in (3.2).

Using the law of total expectation, we have

$$\begin{aligned}
 m_i(1) &= \mathbb{E}_{\mathcal{T}, \mathbf{R}} [\mathbb{E} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}]] \\
 &= \mathbb{E}_{\mathcal{T}, \mathbf{R}} \left[ \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right] \\
 &= \mathbb{E}_{\mathcal{T}} \left[ \hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} \right] \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right], \quad i = 0, 1.
 \end{aligned}$$

Starting from the expression in the last line, we can proceed with the derivation of the DE as we would with the class-conditional mean of the discriminant-averaging RP-LDA infinite ensemble. Note that this is exactly why we end up having  $\bar{m}_i(1) = \bar{m}_i^{M=\infty}$ . Based on the derivation in (Niyazi et al., 2020b), for  $i = 0, 1$ ,

$$\begin{aligned}
 \bar{m}_i(1) &= \frac{1}{2} \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \left[ (-1)^{i+1} \boldsymbol{\mu}^T \left( \frac{p}{n-2} g \boldsymbol{\Sigma} + \mathbf{I}_p \right)^{-1} \boldsymbol{\mu} + \right. \\
 &\quad \left. \left( \frac{1}{n_0} - \frac{1}{n_1} \right) \text{tr} \left\{ \boldsymbol{\Sigma} \left( \frac{p}{n-2} g \boldsymbol{\Sigma} + \mathbf{I}_p \right)^{-1} \right\} \right] + \ln \frac{\pi_1}{\pi_0},
 \end{aligned}$$

where  $g$  satisfies the system of equations defined by

$$\frac{p}{n-2}g = \frac{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)}{1 + \tilde{g}} \quad (\text{A.1})$$

and

$$\tilde{g} = \frac{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)}{n-2} \text{tr} \left\{ \Sigma \left( \mathbf{I}_p + \frac{p}{n-2}g\Sigma \right)^{-1} \right\}, \quad (\text{A.2})$$

and

$$\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) = \frac{\frac{p}{n-2}y^*}{1 - \frac{p}{n-2}y^* \frac{1}{n-2} \text{tr} \left\{ \Sigma \left( \mathbf{I}_p + \frac{p}{n-2}y^*\Sigma \right)^{-1} \right\}},$$

where  $y^*$  is the unique root of the function

$$h(y) = 1 - \frac{p}{d} + \frac{1}{d} \text{tr} \left\{ \left( \mathbf{I}_p + \frac{p}{n-2}y\mathbf{D}_\Sigma \right)^{-1} \right\}$$

which exists when  $p > d$ . Since  $\bar{m}_i(1) = \bar{m}_i^{M=\infty}$ , we denote both DEs by  $\bar{m}_i$ .

### A.1.2 Variance

In this section, we derive the DE of the quantity  $\sigma^2(1)$  defined in (3.3). By making use of the law of total variance with conditioning on the training data and projections (which are independent of  $\mathbf{x}$ ), we have

$$\sigma^2(1) = \mathbb{E}_{\mathcal{T}, \mathbf{R}} [\text{Var} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}]] + \text{Var}_{\mathcal{T}, \mathbf{R}} [\mathbb{E} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}]] \quad (\text{A.3})$$

The second term tends almost-surely to zero, as it is decaying. This can be shown using Lemma 3.1 in (Hachem et al., 2013).

Now, based on the data assumptions on  $\mathbf{x}$ , the inner term of the first term in (A.3) is exactly

$$\text{Var} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}] = \hat{\boldsymbol{\mu}}^T \mathbf{R}^T (\mathbf{R} \hat{\Sigma} \mathbf{R}^T)^{-1} \mathbf{R} \Sigma \mathbf{R}^T (\mathbf{R} \hat{\Sigma} \mathbf{R}^T)^{-1} \mathbf{R} \hat{\boldsymbol{\mu}}. \quad (\text{A.4})$$

We find the DE of (A.4) in what follows. The first term in (A.3) is then the

expectation of this DE by the Vitali convergence theorem, since it can be shown that (A.4) is a uniformly integrable sequence of random variables. This class of random variables have the property that for a sequence  $X_n$  such that  $X_n \asymp X$ , we also have  $\mathbb{E}[X_n] \asymp \mathbb{E}[X]$ .

Note that the rank of the  $p \times p$  matrix  $\hat{\Sigma}$  is at most  $\min\{p, n-2\}$ . Therefore,  $\hat{\Sigma}$  is singular when  $p > n-2$ . Let  $r = \text{rank}(\hat{\Sigma})$ . Then  $\hat{\Sigma} = \mathbf{U}\mathbf{D}\mathbf{U}^T = \mathbf{U}_r\mathbf{D}_r\mathbf{U}_r^T$ , where  $\mathbf{U}_r \in \mathbb{R}^{p \times r}$  contains the  $r$  eigenvectors of  $\hat{\Sigma}$  corresponding to non-zero eigenvalues and  $\mathbf{D}_r \in \mathbb{R}^{r \times r}$  contains the non-zero eigenvalues of  $\hat{\Sigma}$  along its diagonal. This is the compact form of  $\hat{\Sigma}$ . Note that since  $\hat{\Sigma}$  is symmetric (and thus a normal matrix), its pseudoinverse is  $\hat{\Sigma}^+ = \mathbf{U}_r\mathbf{D}_r^{-1}\mathbf{U}_r^T$ . This is made use of later in the derivation. Also note that since we are deriving a DE, access to the actual value of  $r$  is forbidden as it depends on the sample covariance matrix. Nonetheless, we can make use of the fact that under the Gaussian assumptions,  $r = \min\{p, n-2\}$  almost-surely. Keep in mind that  $\mathbf{U}_r^T\mathbf{U}_r = \mathbf{I}_r$ , while, in general,  $\mathbf{U}_r\mathbf{U}_r^T \neq \mathbf{I}_p$ , except when  $r = p$ , i.e.,  $p \leq n-2$ .

We can decompose  $\mathbf{U}$  as  $\mathbf{U} = [\mathbf{U}_r \tilde{\mathbf{U}}_r]$ , where  $\tilde{\mathbf{U}}_r \in \mathbb{R}^{p \times (p-r)}$  has as its columns the eigenvectors corresponding to the zero eigenvalues of  $\hat{\Sigma}$ . Then  $\mathbf{I}_p = \mathbf{U}\mathbf{U}^T = \mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T$ . Let  $\mathbf{R}_r := \mathbf{R}\mathbf{U}_r \in \mathbb{R}^{d \times r}$  and  $\tilde{\mathbf{R}}_r := \mathbf{R}\tilde{\mathbf{U}}_r \in \mathbb{R}^{d \times (p-r)}$ . Define the resolvent  $\mathbf{Q}_2(\beta)$  as

$$\begin{aligned} \mathbf{Q}_2(\beta) &:= (\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \beta\mathbf{I}_d)^{-1} \\ &= (\mathbf{R}_r\mathbf{D}_r\mathbf{R}_r^T + \beta\mathbf{I}_d)^{-1}, \end{aligned}$$

then

$$\begin{aligned} \mathbf{A}(\beta) &:= \mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \beta\mathbf{I}_d)^{-1}\mathbf{R}\Sigma\mathbf{R}^T(\mathbf{R}\hat{\Sigma}\mathbf{R}^T + \beta\mathbf{I}_d)^{-1}\mathbf{R} \\ &= \mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R} \left( \mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T \right) \Sigma \left( \mathbf{U}_r\mathbf{U}_r^T + \tilde{\mathbf{U}}_r\tilde{\mathbf{U}}_r^T \right) \mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R} \\ &= \mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\Sigma\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R} + \mathbf{R}^T\mathbf{Q}_2(\beta)\mathbf{R}_r\mathbf{U}_r^T\Sigma\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R} \\ &\quad + \mathbf{R}^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\Sigma\mathbf{U}_r\mathbf{R}_r^T\mathbf{Q}_2(\beta)\mathbf{R} + \mathbf{R}^T\mathbf{Q}_2(\beta)\tilde{\mathbf{R}}_r\tilde{\mathbf{U}}_r^T\Sigma\tilde{\mathbf{U}}_r\tilde{\mathbf{R}}_r^T\mathbf{Q}_2(\beta)\mathbf{R}. \end{aligned}$$

Overall,

$$\text{Var} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}] = \lim_{\beta \rightarrow 0} \hat{\boldsymbol{\mu}}^T \mathbf{A}(\beta) \hat{\boldsymbol{\mu}}.$$

Now we find the DE of the term  $\hat{\boldsymbol{\mu}}^T \mathbf{A}(\beta) \hat{\boldsymbol{\mu}}$  after which we take the limit as  $\beta \rightarrow 0$ .

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \mathbf{A}(\beta) \hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T \left( \mathbf{U}_r \mathbf{U}_r^T + \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \right) \mathbf{A}(\beta) \left( \mathbf{U}_r \mathbf{U}_r^T + \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \right) \hat{\boldsymbol{\mu}} \\ &= \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{U}_r^T \mathbf{A}(\beta) \mathbf{U}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{U}_r^T \mathbf{A}(\beta) \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \\ &\quad + \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \mathbf{A}(\beta) \mathbf{U}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \mathbf{A}(\beta) \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}}. \end{aligned} \quad (\text{A.5})$$

We consider each term in (A.5) one-by-one. The derivations which follow use the fact that the odd moments of a zero-mean Gaussian random variable are zero. This yields asymptotic simplifications when taking the expectation with respect to  $\tilde{\mathbf{R}}_r$  which is independent of  $\mathbf{R}_r$  and never appears in a resolvent.

For the first term in (A.5), we have

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{U}_r^T \mathbf{A}(\beta) \mathbf{U}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\ &\quad + \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\ &\quad + \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\ &\quad + \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\ &\asymp \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \quad (\text{A.6}) \\ &\quad + \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \quad (\text{A.7}) \end{aligned}$$

For the second term in (A.5), we have



$\mathbf{Q}_1(\beta) = \left( \mathbf{D}_r^{1/2} \mathbf{R}_r^T \mathbf{R}_r \mathbf{D}_r^{1/2} + \beta \mathbf{I}_r \right)^{-1}$ . For (A.6), we have

$$\begin{aligned}
& \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&= \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{D}_r^{-1/2} \mathbf{D}_r^{1/2} \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{D}_r^{1/2} \mathbf{D}_r^{-1/2} \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{D}_r^{-1/2} \mathbf{D}_r^{1/2} \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{D}_r^{1/2} \mathbf{D}_r^{-1/2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&= \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^+ \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^+ \hat{\boldsymbol{\mu}} - 2\beta \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^+ \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{D}_r^{-1/2} \mathbf{Q}_1(\beta) \mathbf{D}_r^{-1/2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&\quad + \beta^2 \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{D}_r^{-1/2} \mathbf{Q}_1(\beta) \mathbf{D}_r^{-1/2} \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{D}_r^{-1/2} \mathbf{Q}_1(\beta) \mathbf{D}_r^{-1/2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}},
\end{aligned}$$

where the second-to-last line makes use of the following relation obtained from the matrix-inversion lemma:

$$\mathbf{D}_r^{1/2} \mathbf{R}_r^T (\mathbf{R}_r \mathbf{D}_r^{1/2} \mathbf{D}_r^{1/2} \mathbf{R}_r^T + \beta \mathbf{I}_d)^{-1} \mathbf{R}_r \mathbf{D}_r^{1/2} = \beta \left[ \frac{1}{\beta} \mathbf{I}_r - \left( \mathbf{D}_r^{1/2} \mathbf{R}_r^T \mathbf{R}_r \mathbf{D}_r^{1/2} + \beta \mathbf{I}_r \right)^{-1} \right].$$

From (Kammoun et al., 2019), we have

$$\mathbf{Q}_1(\beta) \leftrightarrow \mathbf{T}_1(\beta),$$

where

$$\begin{aligned}
\mathbf{T}_1(\beta) &= \frac{1}{\beta} (\mathbf{I}_r + \tilde{\nu}_1(\beta) \mathbf{D}_r)^{-1} \\
\tilde{\mathbf{T}}_1(\beta) &= \frac{1}{\beta (1 + \frac{r}{d} \nu_1(\beta))} \mathbf{I}_d
\end{aligned}$$

and

$$\begin{aligned}
\nu_1(\beta) &= \frac{1}{\beta} \frac{1}{r} \text{tr} \{ \mathbf{D}_r (\mathbf{I}_r + \tilde{\nu}_1(\beta) \mathbf{D}_r)^{-1} \} \\
\tilde{\nu}_1(\beta) &= \frac{1}{\beta (1 + \frac{r}{d} \nu_1(\beta))}.
\end{aligned} \tag{A.14}$$

Using the above relations, we have

$$\begin{aligned} & \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\ & \asymp \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-1} \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-1} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\ & \quad + \left( \frac{1}{\tilde{\nu}_1(\beta)} \right)^2 \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \left( \frac{\beta^2 \theta(\mathbf{C}) \tilde{\theta}}{1 - \beta^2 \theta(\mathbf{D}_r) \tilde{\theta}} \right) \end{aligned}$$

where

$$\begin{aligned} \beta^2 \theta(\mathbf{D}_r) &= \frac{\beta^2}{r} \text{tr} \{ \mathbf{D}_r \mathbf{T}_1(\beta) \mathbf{D}_r \mathbf{T}_1(\beta) \} \\ &= \frac{1}{(\tilde{\nu}_1(\beta))^2} \frac{1}{r} \text{tr} \left\{ \mathbf{D} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \mathbf{D} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} \end{aligned}$$

$$\begin{aligned} \beta^2 \theta(\mathbf{C}) &= \frac{\beta^2}{r} \text{tr} \{ \mathbf{D}_r \mathbf{T}_1(\beta) \mathbf{C} \mathbf{T}_1(\beta) \} \\ &= \frac{1}{r} \text{tr} \{ \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \} - \frac{2}{r} \text{tr} \left\{ \mathbf{U}^T \boldsymbol{\Sigma} \mathbf{U} \mathbf{D} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} \\ & \quad + \frac{1}{r} \text{tr} \left\{ \mathbf{D} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \mathbf{U}^T \boldsymbol{\Sigma} \mathbf{U} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \mathbf{D} \right\} \\ &= \frac{1}{(\tilde{\nu}_1(\beta))^2} \frac{1}{r} \text{tr} \left\{ \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \mathbf{U}^T \boldsymbol{\Sigma} \mathbf{U} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} - \frac{1}{r} \text{tr} \{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \}, \end{aligned} \tag{A.15}$$

and

$$\tilde{\theta} = \frac{r}{d} \left( \frac{1}{\beta \left( 1 + \frac{r}{d} \nu_1(\beta) \right)} \right)^2 = \frac{r}{d} (\tilde{\nu}_1(\beta))^2,$$

where  $\mathbf{C} := \mathbf{D}_r^{-1/2} \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{D}_r^{-1/2}$ .

Now consider (A.7). Let  $\mathbf{a} = \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}}$  and  $\tilde{\mathbf{R}}_r$  have rows  $\tilde{\mathbf{r}}_1, \dots, \tilde{\mathbf{r}}_d$ . We

have the intermediate convergence

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} &= \sum_{i,j} a_i a_j \tilde{\mathbf{r}}_i^T \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{r}}_j \\
&\asymp \sum_i a_i^2 \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \\
&= \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2^2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}}
\end{aligned}$$

We can show that

$$\hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2^2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \asymp \frac{1}{d} \text{tr} \left\{ \mathbf{Q}_2^2(\beta) \right\} \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \frac{1}{d} \text{tr} \left\{ \mathbf{Q}_2(\beta) \right\} \mathbf{D}_r + \mathbf{I}_r \right)^{-2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}}.$$

Using (Kammoun et al., 2019), we have

$$\mathbf{Q}_2(\beta) \leftrightarrow \mathbf{T}_2(\beta) \tag{A.16}$$

where

$$\begin{aligned}
\mathbf{T}_2(\beta) &= \frac{1}{\beta (1 + \tilde{\nu}_2(\beta))} \mathbf{I}_d \\
\tilde{\mathbf{T}}_2(\beta) &= \frac{1}{\beta} (\mathbf{I}_r + \nu_2(\beta) \mathbf{D}_r)^{-1}
\end{aligned}$$

and

$$\begin{aligned}
\nu_2(\beta) &= \frac{1}{\beta (1 + \tilde{\nu}_2(\beta))} \\
\tilde{\nu}_2(\beta) &= \frac{1}{\beta} \frac{1}{d} \text{tr} \left\{ \mathbf{D}_r (\mathbf{I}_r + \nu_2(\beta) \mathbf{D}_r)^{-1} \right\}.
\end{aligned} \tag{A.17}$$

From (A.16), we have

$$\frac{1}{d} \text{tr} \left\{ \mathbf{Q}_2(\beta) \right\} \asymp \frac{1}{d} \text{tr} \left\{ \mathbf{T}_2(\beta) \right\}$$

Now, let's find the DE of  $\frac{1}{d} \text{tr} \left\{ \mathbf{Q}_2^2(\beta) \right\}$ . First, using the systems of equations

in (A.14) and (A.17), we can show that  $\nu_2(\beta) = \tilde{\nu}_1(\beta)$ . Since

$$\frac{1}{d} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} = - \frac{d \left[ \frac{1}{d} \text{tr} \{ \mathbf{Q}_2(\beta) \} \right]}{d\beta},$$

then

$$\frac{1}{d} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} \asymp - \frac{d \left[ \frac{1}{d} \text{tr} \{ \mathbf{T}_2(\beta) \} \right]}{d\beta},$$

i.e., the limit of the derivative is the derivative of the limit. To justify this, first note that  $\frac{1}{d} \text{tr} \{ \mathbf{Q}_2(\beta) \}$  is a Stieltjes transform which is analytic outside the support of the spectrum of  $\mathbf{R} \hat{\Sigma} \mathbf{R}^T$ . Since the support of the spectrum is bounded away from zero, taking  $\beta \rightarrow 0$  ensures that  $\frac{1}{d} \text{tr} \{ \mathbf{Q}_2(\beta) \}$  is analytic. Similarly,  $\frac{1}{d} \text{tr} \{ \mathbf{T}_2(\beta) \}$  is a Stieltjes transform which is analytic outside the support of the limiting spectrum of  $\mathbf{R} \hat{\Sigma} \mathbf{R}^T$ , and since the support of the limiting spectrum is bounded away from zero, taking  $\beta \rightarrow 0$  ensures that the  $\frac{1}{d} \text{tr} \{ \mathbf{T}_2(\beta) \}$  is analytic. Since both  $\frac{1}{d} \text{tr} \{ \mathbf{Q}_2(\beta) \}$  and its limit  $\frac{1}{d} \text{tr} \{ \mathbf{T}_2(\beta) \}$  are analytic, it follows that all derivatives of  $\frac{1}{d} \text{tr} \{ \mathbf{Q}_2(\beta) \}$  of any order converge to the corresponding derivatives of  $\frac{1}{d} \text{tr} \{ \mathbf{T}_2(\beta) \}$ . Then, because

$$\begin{aligned} \frac{1}{d} \text{tr} \{ \mathbf{T}_2(\beta) \} &= \frac{1}{d} \text{tr} \left\{ \frac{1}{\beta(1 + \tilde{\nu}_2(\beta))} \right\} \\ &= \nu_2(\beta) \\ &= \tilde{\nu}_1(\beta), \end{aligned}$$

we have

$$\frac{1}{d} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} \asymp -\tilde{\nu}'_1(\beta).$$

We can solve the following set of equations (obtained by differentiating the system

of equations (A.14) for  $\tilde{\nu}'_1(\beta)$ :

$$\begin{aligned}
\nu'_1(\beta) &= -\frac{\tilde{\nu}'_1(\beta)}{\beta} \frac{1}{r} \text{tr} \left\{ \mathbf{D}_r (\mathbf{I}_r + \tilde{\nu}_1(\beta) \mathbf{D}_r)^{-1} \mathbf{D}_r (\mathbf{I}_r + \tilde{\nu}_1(\beta) \mathbf{D}_r)^{-1} \right\} \\
&\quad - \frac{1}{\beta^2} \frac{1}{r} \text{tr} \left\{ \mathbf{D}_r (\mathbf{I}_r + \tilde{\nu}_1(\beta) \mathbf{D}_r)^{-1} \right\} \\
&= -\tilde{\nu}'_1(\beta) \beta \theta(\mathbf{D}_r) - \frac{1}{\beta} \nu_1(\beta) \\
\tilde{\nu}'_1(\beta) &= -\frac{r}{d} \frac{\nu'_1(\beta)}{\beta} \frac{1}{(1 + \frac{r}{d} \nu_1(\beta))^2} - \frac{1}{\beta^2 (1 + \frac{r}{d} \nu_1(\beta))} \\
&= -\nu'_1(\beta) \beta \tilde{\theta} - \frac{1}{\beta} \tilde{\nu}_1(\beta),
\end{aligned}$$

from which

$$\tilde{\nu}'_1(\beta) = \frac{\nu_1(\beta) \tilde{\theta} - \frac{1}{\beta} \tilde{\nu}_1(\beta)}{1 - \beta^2 \theta(\mathbf{D}_r) \tilde{\theta}}.$$

So, overall we have

$$\begin{aligned}
&\hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&\asymp -\frac{\tilde{\nu}'_1(\beta)}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \frac{1}{d} \text{tr} \left\{ \mathbf{T}_2(\beta) \right\} \mathbf{D}_r + \mathbf{I}_r \right)^{-2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&= -\frac{\tilde{\nu}'_1(\beta)}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \mathbf{U}_r (\tilde{\nu}_1(\beta) \mathbf{D}_r + \mathbf{I}_r)^{-2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&= -\frac{\tilde{\nu}'_1(\beta)}{(\tilde{\nu}_1(\beta))^2} \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}}.
\end{aligned}$$

Applying the same techniques to (A.8), we can show that

$$\hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \asymp \tilde{\nu}_1(\beta) \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-1} \mathbf{U}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}}.$$

For (A.9), we have

$$\hat{\boldsymbol{\mu}}^T \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \asymp 0.$$

The terms (A.10) and (A.11) are just the transpose of (A.8) and (A.9). For

(A.12), we have

$$\begin{aligned} & \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \mathbf{R}_r \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \mathbf{R}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \\ & \asymp -\hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \frac{\tilde{\nu}'_1(\beta)}{(\tilde{\nu}_1(\beta))^2} \frac{1}{d} \text{tr} \left\{ \boldsymbol{\Sigma} \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-2} \mathbf{U}_r^T \right\}, \end{aligned}$$

and for the final term (A.13), we have

$$\begin{aligned} & \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \\ & = \sum_{i,j,k,l} \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_j \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{k,l} \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_l^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_j. \end{aligned}$$

It is easy to see that only three cases survive asymptotically in this summation:

1.  $i = j = k = l$
2.  $i = k, j = l, i \neq j$
3.  $i = j, k = l, i \neq k$

For the first case,

$$\begin{aligned} & \sum_{i=j=k=l} \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_j \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{k,l} \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_l^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_j \\ & = \sum_i \left( \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \right)^2 \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{i,i} \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_i \\ & \asymp \left[ \frac{2}{d^2} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} + \left( \frac{1}{d} \text{tr} \{ \mathbf{Q}_2(\beta) \} \right)^2 \right] \sum_i \left( \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \right)^2 \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{i,i} \\ & \asymp (\tilde{\nu}_1(\beta))^2 \sum_i \left( \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \right)^2 \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{i,i}, \end{aligned} \tag{A.18}$$

where the third line uses the expectation of the term  $\tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_i$ . For

the second case,

$$\begin{aligned}
& \sum_{i=k, j=l, i \neq j} \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_j \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{k,l} \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_l^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_j \\
&= \sum_{i \neq j} \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_j \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{i,j} \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_i \tilde{\mathbf{r}}_j^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_j \\
&\asymp \left( \frac{1}{d} \text{tr} \{ \mathbf{Q}_2(\beta) \} \right)^2 \sum_{i \neq j} \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_j \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{i,j} \\
&\asymp (\tilde{\nu}_1(\beta))^2 \sum_{i \neq j} \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_j \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{i,j}. \tag{A.19}
\end{aligned}$$

For the third case, we have

$$\begin{aligned}
& \sum_{i=j, k=l, i \neq k} \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_j \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{k,l} \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_l^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_j \\
&= \sum_{i \neq k} \left( \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \right)^2 \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{k,k} \tilde{\mathbf{r}}_i^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_k \tilde{\mathbf{r}}_k^T \mathbf{Q}_2(\beta) \tilde{\mathbf{r}}_i \\
&\asymp \frac{1}{d^2} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} \sum_{i \neq k} \left( \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \right)^2 \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{k,k} \\
&= \frac{1}{d} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} - \frac{1}{d^2} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} \sum_i \left( \left[ \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \right]_i \right)^2 \left[ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right]_{i,i} \\
&\asymp \frac{1}{d} \text{tr} \{ \mathbf{Q}_2^2(\beta) \} \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \\
&\asymp -\tilde{\nu}'_1(\beta) \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}}. \tag{A.20}
\end{aligned}$$

Combining (A.18), (A.19), and (A.20), we have overall

$$\begin{aligned}
& \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{R}}_r^T \mathbf{Q}_2(\beta) \tilde{\mathbf{R}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \\
&\asymp (\tilde{\nu}_1(\beta))^2 \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} - \tilde{\nu}'_1(\beta) \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}}. \tag{A.21}
\end{aligned}$$

Combining the above derivations starting from (A.5) to (A.21), we have

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^T \mathbf{A}(\beta) \hat{\boldsymbol{\mu}} &\asymp \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-1} \mathbf{U}_r^T \boldsymbol{\Sigma} \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-1} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&+ \left[ \left( \frac{\beta^2 \theta(\mathbf{C}) \tilde{\boldsymbol{\theta}}}{1 - \beta^2 \theta(\mathbf{D}_r) \tilde{\boldsymbol{\theta}}} \right) - \tilde{\nu}'_1(\beta) \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \right] \frac{1}{(\tilde{\nu}_1(\beta))^2} \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-2} \mathbf{U}_r^T \hat{\boldsymbol{\mu}} \\
&+ 2\tilde{\nu}_1(\beta) \hat{\boldsymbol{\mu}}^T \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-1} \mathbf{U}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \\
&- \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} \frac{\tilde{\nu}'_1(\beta)}{(\tilde{\nu}_1(\beta))^2} \frac{1}{d} \text{tr} \left\{ \boldsymbol{\Sigma} \mathbf{U}_r \left( \mathbf{D}_r + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_r \right)^{-2} \mathbf{U}_r^T \right\} \\
&+ (\tilde{\nu}_1(\beta))^2 \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}} - \tilde{\nu}'_1(\beta) \frac{1}{d} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\} \hat{\boldsymbol{\mu}}^T \tilde{\mathbf{U}}_r \tilde{\mathbf{U}}_r^T \hat{\boldsymbol{\mu}}.
\end{aligned}$$

Through a series of manipulations in which we express everything in terms of  $\mathbf{D}$  instead of  $\mathbf{D}_r$  and also by using the relation

$$\tilde{\nu}'_1(\beta) = -\frac{(\tilde{\nu}_1(\beta))^2}{1 - \beta^2 \theta(\mathbf{D}_r) \tilde{\boldsymbol{\theta}}},$$

obtained through the system of equations (A.14) and by expressing (A.15) as

$$\beta^2 \theta(\mathbf{C}) = \beta^2 \theta(\mathbf{C}') - \frac{1}{r} \text{tr} \left\{ \tilde{\mathbf{U}}_r^T \boldsymbol{\Sigma} \tilde{\mathbf{U}}_r \right\}$$

where

$$\beta^2 \theta(\mathbf{C}') = \frac{1}{(\tilde{\nu}_1(\beta))^2} \frac{1}{r} \text{tr} \left\{ \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \mathbf{U}^T \boldsymbol{\Sigma} \mathbf{U} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\},$$

we have the simplification

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^T \mathbf{A}(\beta) \hat{\boldsymbol{\mu}} &\asymp \hat{\boldsymbol{\mu}}^T \mathbf{U} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \mathbf{U}^T \boldsymbol{\Sigma} \mathbf{U} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \mathbf{U}^T \hat{\boldsymbol{\mu}} \\
&+ \left( \frac{\beta^2 \theta(\mathbf{C}') \tilde{\boldsymbol{\theta}}}{1 - \beta^2 \theta(\mathbf{D}_r) \tilde{\boldsymbol{\theta}}} \right) \frac{1}{(\tilde{\nu}_1(\beta))^2} \hat{\boldsymbol{\mu}}^T \mathbf{U} \left( \mathbf{D} + \frac{1}{\tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-2} \mathbf{U}^T \hat{\boldsymbol{\mu}} \quad (\text{A.22})
\end{aligned}$$

Now what must be done is to remove the randomness from the training. This

appears in  $\hat{\boldsymbol{\mu}}$ ,  $\mathbf{D}$ , and in the current definition of  $\tilde{\nu}_1(\beta)$ .

First, we derive  $\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)$  in such a way that it depends only on the true statistics. Using the equations in (A.14), it can be shown that

$$1 - \frac{p}{d} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \frac{1}{d} \text{tr} \left\{ \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} = 0 \quad (\text{A.23})$$

Using the fact that  $\hat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \boldsymbol{\Sigma}^{1/2} \mathbf{Z} \mathbf{Z}^T \boldsymbol{\Sigma}^{1/2}$  for some  $\mathbf{Z} \in \mathbb{R}^{p \times (n-2)}$  with i.i.d. standard Gaussian entries and by eigendecomposing  $\boldsymbol{\Sigma}$  as  $\boldsymbol{\Sigma} = \mathbf{V} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{V}^T$ , we have

$$\begin{aligned} \frac{1}{d} \text{tr} \left\{ \left( \hat{\boldsymbol{\Sigma}} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} &= \frac{1}{d} \text{tr} \left\{ \left( \frac{1}{n-2} \boldsymbol{\Sigma}^{1/2} \mathbf{Z} \mathbf{Z}^T \boldsymbol{\Sigma}^{1/2} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} \\ &\sim \frac{1}{d} \text{tr} \left\{ \left( \frac{1}{n-2} \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2} \mathbf{Z} \mathbf{Z}^T \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} \end{aligned}$$

From (Kammoun et al., 2019), we have

$$\mathbf{W}(\gamma) \leftrightarrow \mathbf{E}(\gamma),$$

where

$$\mathbf{W}(\gamma) = \left( \frac{1}{n-2} \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2} \mathbf{Z} \mathbf{Z}^T \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2} - \gamma \mathbf{I}_p \right)^{-1},$$

$$\mathbf{E}(\gamma) = -\frac{1}{\gamma} \left( \mathbf{I}_p + \frac{p}{n-2} g(\gamma) \mathbf{D}_{\boldsymbol{\Sigma}} \right)^{-1},$$

$$\frac{p}{n-2} g(\gamma) = -\frac{1}{\gamma} \frac{1}{1 + \tilde{g}(\gamma)},$$

and

$$\tilde{g}(\gamma) = -\frac{1}{\gamma} \frac{1}{p} \text{tr} \left\{ \frac{p}{n-2} \mathbf{D}_{\boldsymbol{\Sigma}} \left( \mathbf{I}_p + \frac{p}{n-2} g(\gamma) \mathbf{D}_{\boldsymbol{\Sigma}} \right)^{-1} \right\}$$

from which it follows that

$$\frac{1}{d} \text{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \mathbf{I}_p \right)^{-1} \right\} \asymp \frac{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)}{d} \text{tr} \left\{ \left( \mathbf{I}_p + \frac{p}{n-2} g \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \mathbf{D}_{\Sigma} \right)^{-1} \right\}$$

Let  $g := g \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)$  and  $\tilde{g} := \tilde{g} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)$ . We now have

$$1 - \frac{p}{d} + \frac{1}{d} \text{tr} \left\{ \left( \mathbf{I}_p + \frac{p}{n-2} g \mathbf{D}_{\Sigma} \right)^{-1} \right\} \asymp 0. \quad (\text{A.24})$$

Using (A.24), the quantity  $g$  can be solved for as the unique root  $y^*$  of the monotonically decreasing function

$$\begin{aligned} h(y) &= 1 - \frac{p}{d} + \frac{1}{d} \text{tr} \left\{ \left( \mathbf{I}_p + \frac{p}{n-2} y \mathbf{D}_{\Sigma} \right)^{-1} \right\} \\ &= 1 - \frac{p}{d} + \frac{1}{d} \sum_{i=1}^p \frac{1}{1 + \frac{p}{n-2} \lambda_i(\Sigma) y} \end{aligned}$$

which exists when  $p > d$ . It can be shown that  $g \asymp y^*$ . Then combining (A.1) and (A.2), we can solve for  $\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)$  in terms of  $g$ , and so we have

$$\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) = \frac{\frac{p}{n-2} y^*}{1 - \frac{p}{n-2} y^* \frac{1}{n-2} \text{tr} \left\{ \mathbf{D}_{\Sigma} \left( \mathbf{I}_p + \frac{p}{n-2} y^* \mathbf{D}_{\Sigma} \right)^{-1} \right\}}.$$

By dealing with the randomness from the sample covariance in (A.22) using

similar techniques, followed by taking the limit as  $\beta \rightarrow 0$ , we obtain

$$\begin{aligned} \lim_{\beta \rightarrow 0} \hat{\boldsymbol{\mu}}^T \mathbf{A}(\beta) \hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T \mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T)^{-1} \mathbf{R} \boldsymbol{\Sigma} \mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T)^{-1} \mathbf{R} \hat{\boldsymbol{\mu}} \\ &\asymp \left( \frac{1}{1-\Omega} \right) \hat{\boldsymbol{\mu}}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{V}^T \hat{\boldsymbol{\mu}} + \\ &\quad \frac{1}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \right)^2} \left[ \hat{\boldsymbol{\mu}}^T \mathbf{V} \mathbf{E}^2 \mathbf{V}^T \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{V}^T \hat{\boldsymbol{\mu}} \left( \frac{\left( \frac{\frac{p}{n-2} g}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \text{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2}{1-\Omega} \right) \right] \times \\ &\quad \left( \frac{\left( \frac{1}{1-\Omega} \right) \frac{1}{d} \text{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2}{1 - \frac{p}{d} + \frac{2}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \frac{1}{d} \text{tr} \mathbf{E} - \frac{1}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \right)^2} \left[ \frac{1}{d} \text{tr} \mathbf{E}^2 + \frac{1}{d} \text{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2 \left( \frac{\left( \frac{\frac{p}{n-2} g}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \text{tr} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E}^2}{1-\Omega} \right) \right]} \right), \end{aligned}$$

where

$$\Omega = \left( \frac{\frac{p}{n-2} g}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \text{tr} \{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \}.$$

The final step is to remove the randomness coming from the sample means in  $\hat{\boldsymbol{\mu}}$ .

Using

$$\hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1} - \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0},$$

where  $\mathbf{Z}_i \in \mathbb{R}^{p \times n_i}$ ,  $i = 0, 1$  has i.i.d.  $\mathcal{N}(0, 1)$  entries. Taking the expectation over  $\mathbf{Z}_i \mathbf{1}$ ,  $i = 0, 1$ , while making use of the fact that  $\frac{\mathbf{Z}_i \mathbf{1}}{n_i} \sim \mathcal{N}\left(\mathbf{0}_p, \frac{1}{n_i} \mathbf{I}_p\right)$ ,  $i = 0, 1$ , we

have

$$\begin{aligned}
\bar{\sigma}^2(1) &= \left( \frac{1}{1-\Omega} \right) \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\Sigma} \mathbf{E} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_{\Sigma} \mathbf{E} \mathbf{D}_{\Sigma} \mathbf{E} \} \right] + \\
&\quad \frac{1}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \right)^2} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{E}^2 \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2 + \right. \\
&\quad \left. \left( \boldsymbol{\mu}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\Sigma} \mathbf{E} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \mathbf{D}_{\Sigma} \mathbf{E} \mathbf{D}_{\Sigma} \mathbf{E} \right) \left( \frac{\left( \frac{\frac{p}{n-2} g}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2}{1-\Omega} \right) \right] \times \\
&\quad \left( \frac{\left( \frac{1}{1-\Omega} \right) \frac{1}{d} \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2}{1 - \frac{p}{d} + \frac{2}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \frac{1}{d} \text{tr} \mathbf{E} - \frac{1}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \right)^2} \left[ \frac{1}{d} \text{tr} \mathbf{E}^2 + \frac{1}{d} \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2 \left( \frac{\left( \frac{\frac{p}{n-2} g}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)^2 \frac{1}{n-2} \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2}{1-\Omega} \right) \right]} \right).
\end{aligned} \tag{A.25}$$

## Proof that the single RP-LDA discriminant variance is asymptotically greater than that of the infinite ensemble

We simply prove that  $\bar{\sigma}^2(1) > \bar{\sigma}_{M=\infty}^2$ . Using the expressions in (A.25) and (A.29), we have

$$\begin{aligned} \bar{\sigma}^2(1) - \bar{\sigma}_{M=\infty}^2 &= \frac{1}{\left(\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)\right)^2} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{E}^2 \mathbf{V}^T \boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1}\right) \text{tr} \{ \mathbf{D}_{\Sigma} \mathbf{E}^2 \} + \right. \\ &\quad \left. \left( \boldsymbol{\mu}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\Sigma} \mathbf{E} \mathbf{V}^T \boldsymbol{\mu} + \left(\frac{1}{n_0} + \frac{1}{n_1}\right) \text{tr} \{ \mathbf{D}_{\Sigma} \mathbf{E} \mathbf{D}_{\Sigma} \mathbf{E} \} \right) \left( \frac{\left(\frac{\frac{p}{n-2}g}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)}\right)^2 \frac{1}{n-2} \text{tr} \{ \mathbf{D}_{\Sigma} \mathbf{E}^2 \}}{1 - \Omega} \right) \right] \times \\ &\quad \left( \frac{\left(\frac{1}{1-\Omega}\right) \frac{1}{d} \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2}{1 - \frac{p}{d} + \frac{2}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \frac{1}{d} \text{tr} \mathbf{E} - \frac{1}{\left(\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)\right)^2} \left[ \frac{1}{d} \text{tr} \mathbf{E}^2 + \frac{1}{d} \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2 \left( \frac{\left(\frac{\frac{p}{n-2}g}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)}\right)^2 \frac{1}{n-2} \text{tr} \mathbf{D}_{\Sigma} \mathbf{E}^2}{1 - \Omega} \right) \right]} \right) \end{aligned} \quad (\text{A.26})$$

Now we must show that each of the constituent terms of (A.26) is positive. The term  $1 - \Omega$  fits the form of the term  $1 - t^2 \gamma_n(t) \tilde{\gamma}_n(t)$  in the paper (Hachem et al., 2008) in which it was shown to be positive. Additionally, all traces and quadratic terms in the first and second lines of (A.26) are positive since the matrices involved are positive definite. What remains is the denominator of the fraction in the last line. This term comes from taking the asymptotic limit of the

term  $1 - \lim_{\beta \rightarrow 0} \beta^2 \theta(\mathbf{D}_r) \tilde{\theta}$  which can be expressed as

$$1 - \lim_{\beta \rightarrow 0} \beta^2 \theta(\mathbf{D}_r) \tilde{\theta} = 1 - \frac{p}{d} + \frac{2}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \frac{1}{d} \text{tr} \left\{ \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \right\} \\ - \frac{1}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \right)^2} \frac{1}{d} \text{tr} \left\{ \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \right\} \quad (\text{A.27})$$

Using (A.23), equation (A.27) simplifies to

$$1 - \lim_{\beta \rightarrow 0} \beta^2 \theta(\mathbf{D}_r) \tilde{\theta} = \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \frac{1}{d} \text{tr} \left\{ \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \right\} \\ - \frac{1}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \right)^2} \frac{1}{d} \text{tr} \left\{ \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \right\}.$$

Let  $\mathbf{G} := \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \mathbf{D} + \mathbf{I}_p$ . Then, using the relation  $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$

with  $\mathbf{A}^{-1} = \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)$  and  $\mathbf{B}^{-1} = \frac{1}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) \right)^2} \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right) \mathbf{W} \left( -\frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta)} \right)$ ,

we have

$$1 - \lim_{\beta \rightarrow 0} \beta^2 \theta(\mathbf{D}_r) \tilde{\theta} = \frac{1}{d} \text{tr} \{ \mathbf{G}^{-1} (\mathbf{G}^2 - \mathbf{G}) \mathbf{G}^{-2} \} \\ = \frac{1}{d} \sum_{i=1}^p \frac{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) d_i + 1 \right)^2 - \left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) d_i + 1 \right)}{\left( \lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) d_i + 1 \right)^3} \\ > 0,$$

since  $\lim_{\beta \rightarrow 0} \tilde{\nu}_1(\beta) d_i + 1 > 1$ , where  $d_i$  is the  $i^{\text{th}}$  entry along the diagonal of the diagonal matrix  $\mathbf{D}$ .

## A.2 Discriminant-averaging RP-LDA finite ensemble class-conditional discriminant statistics

### A.2.1 Means

In this section, we derive the DE of the quantity  $m_i(M)$ ,  $i = 0, 1$ , defined in (3.5). By the law of total expectation, we have

$$\begin{aligned}
m_i(M) &= \mathbb{E}_{\mathcal{T}, \mathbf{R}} [\mathbb{E} [W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}]] \\
&= \frac{1}{M} \sum_{k=1}^M \mathbb{E}_{\mathcal{T}, \mathbf{R}} [\mathbb{E} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}]] \\
&\asymp \frac{1}{M} \sum_{k=1}^M \bar{m}_i(1) \\
&= \bar{m}_i(1), \quad i = 0, 1,
\end{aligned}$$

where the convergence in the second-to-last line is proven in Appendix A.1.1. Thus,  $\bar{m}_i(M) = \bar{m}_i(1)$  and we denote both DEs by  $\bar{m}_i$ .

### A.2.2 Variance

In this section, we derive the DE of the quantity  $\sigma^2(M)$ , defined in (3.6). By the law of total variance,

$$\begin{aligned}
\sigma^2(M) &= \mathbb{E}_{\mathcal{T}, \mathbf{R}} [\text{Var} [W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}]] \\
&\quad + \text{Var}_{\mathcal{T}, \mathbf{R}} [\mathbb{E} [W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}]]. \quad (\text{A.28})
\end{aligned}$$

For a similar reason to that in Appendix A.1.2, the second term in (A.28) is asymptotically zero. Considering the inner term of the first term, we have

$$\begin{aligned}
& \text{Var} \left[ W_{\text{disc-avg}} \left( \mathbf{x}, \{\mathbf{R}_k\}_{k=1}^M \right) \middle| \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M \right] \\
&= \text{Var} \left[ \frac{1}{M} \sum_{k=1}^M W_{\text{RP-LDA}} \left( \mathbf{x}, \mathbf{R}_k \right) \middle| \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M \right] \\
&= \frac{1}{M^2} \sum_{k=1}^M \text{Var} [W_{\text{RP-LDA}} \left( \mathbf{x}, \mathbf{R}_k \right) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}_k] \\
&\quad + \frac{1}{M^2} \sum_{k \neq j}^M \text{Cov} [W_{\text{RP-LDA}} \left( \mathbf{x}, \mathbf{R}_k \right), W_{\text{RP-LDA}} \left( \mathbf{x}, \mathbf{R}_j \right) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}_k, \mathbf{R}_j] \\
&\asymp \frac{1}{M} \bar{\sigma}^2(1) + \frac{M-1}{M^2} \bar{\sigma}_{M=\infty}^2 \\
&= \frac{1}{M} \bar{\sigma}^2(1) + \left( 1 - \frac{1}{M} \right) \bar{\sigma}_{M=\infty}^2
\end{aligned}$$

where the convergence in the second-to-last line follows from the proof in Appendix A.1.2 and also the fact that

$$\begin{aligned}
& \text{Cov} [W_{\text{RP-LDA}} \left( \mathbf{x}, \mathbf{R}_k \right), W_{\text{RP-LDA}} \left( \mathbf{x}, \mathbf{R}_j \right) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}] \\
&= \hat{\boldsymbol{\mu}}^T \mathbf{R}_k^T (\mathbf{R}_k \hat{\boldsymbol{\Sigma}} \mathbf{R}_k^T)^{-1} \mathbf{R}_k \boldsymbol{\Sigma} \mathbf{R}_j^T (\mathbf{R}_j \hat{\boldsymbol{\Sigma}} \mathbf{R}_j^T)^{-1} \mathbf{R}_j \hat{\boldsymbol{\mu}} \\
&\asymp \hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T)^{-1} \mathbf{R} \right] \boldsymbol{\Sigma} \mathbb{E}_{\mathbf{R}} \left[ \mathbf{R}^T (\mathbf{R} \hat{\boldsymbol{\Sigma}} \mathbf{R}^T)^{-1} \mathbf{R} \right] \hat{\boldsymbol{\mu}} \\
&\asymp \bar{\sigma}_{M=\infty}^2.
\end{aligned}$$

The exact expression of  $\bar{\sigma}_{M=\infty}^2$  is derived in (Niyazi et al., 2020b) as

$$\bar{\sigma}_{M=\infty}^2 = \frac{1}{1-\Omega} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{V} \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{E} \} \right]. \quad (\text{A.29})$$

### A.3 RP-LDA ensemble classifier error analysis

This section provides DEs for the probabilities of misclassification of the discriminant-averaging, MAP, and vote-averaging RP-LDA ensemble classifiers.

### A.3.1 Discriminant-averaging RP-LDA ensemble classifier

The expected probability of misclassification DE of the discriminant-averaging RP-LDA ensemble classifier composed of  $M$  RP-LDA discriminants is

$$\pi_0 \Phi \left( \frac{\bar{m}_0}{\sqrt{\bar{\sigma}^2(M)}} \right) + \pi_1 \Phi \left( -\frac{\bar{m}_1}{\sqrt{\bar{\sigma}^2(M)}} \right), \quad (\text{A.30})$$

where  $M = 1, 2, \dots$ . This statement claims the convergence of the expected probability of misclassification of the discriminant-averaging RP-LDA ensemble (over training and projections) to the probability of misclassification computed using the distribution of the asymptotic discriminant stated in Theorem 4.4.2. This follows from the convergence in distribution in Theorem 4.4.2 and Lemma 2.11 in (Van der Vaart, 2000). Note that the convergence is not in the probabilistic sense; the probability of misclassification is conditioned on the training and random projections before applying Lemma 2.11 to obtain its limit. The limit of the expected probability of misclassification over the training and random projections (A.30) is then simply the expectation of the first limit. This follows by the bounded convergence theorem since the probability measure is upper bounded by 1.

### A.3.2 MAP RP-LDA ensemble classifier

The expected probability of misclassification DE of the MAP RP-LDA ensemble classifier composed of  $M$  RP-LDA discriminants is

$$\pi_0 \Phi \left( \frac{-\frac{1}{2} \frac{(\bar{m}_1 - \bar{m}_0)^2}{\bar{\sigma}^2(M)} + \ln \frac{\pi_1}{\pi_0}}{\sqrt{\frac{(\bar{m}_1 - \bar{m}_0)^2}{\bar{\sigma}^2(M)}}} \right) + \pi_1 \Phi \left( \frac{-\frac{1}{2} \frac{(\bar{m}_1 - \bar{m}_0)^2}{\bar{\sigma}^2(M)} - \ln \frac{\pi_1}{\pi_0}}{\sqrt{\frac{(\bar{m}_1 - \bar{m}_0)^2}{\bar{\sigma}^2(M)}}} \right),$$

where  $M = 1, 2, \dots$ . This statement claims the convergence of the expected probability of misclassification of the MAP RP-LDA ensemble (over training and projections) to the probability of misclassification computed using the distribu-

tion of the asymptotic discriminant, which can be derived easily from Theorem 4.4.2. This result then follows from this convergence in distribution and Lemma 2.11 in (Van der Vaart, 2000). Again, the convergence is not in the probabilistic sense.

### A.3.3 Vote-averaging RP-LDA ensemble classifier

As stated in Theorem 3.4.4, the asymptotic distribution of the vote-averaging RP-LDA ensemble class-conditional discriminant times  $M$  is a correlated binomial having  $M$  trials, probability of success  $\bar{p}_i$ ,  $i = 0, 1$ , and constant correlation  $\bar{\rho}_i$  between trial outcomes. In this case, however, knowing the distribution is not enough to determine the asymptotic probability of misclassification. This is because the *correlated* binomial Probability Mass Function (PMF) is not uniquely specified by the correlation coefficient(s) and probability of success. Additional information pertaining to the conditional correlations is needed. This is the reason why there are various models for correlated binomials based on different assumptions.

One of these models is Moody's model (Witt, 2004). Moody's correlated binomial model makes the assumption that the conditional correlations of the outcomes of any two trials given that any subset of the others are all successes is constant. To test this model, we generate the empirical PMF of the vote-averaging RP-LDA ensemble classifier. This model fits the empirical PMF well, at least up to  $M = 35$ , beyond which numerical issues occur which hinder the accurate computation of Moody's PMF. This seems to suggest that the constant conditional correlation condition holds for our setup.

Through our own numerical investigation, we find that this condition, in fact, does not hold, although the conditional correlations are close enough that the corresponding conditional probabilities of success end up being very close to those predicted by Moody's model. Since Moody's PMF is characterized by these conditional probabilities, this might explain why we have a close match.

Denote by  $p_i(k)$  the probability that  $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i$  asymptotically takes the value  $k$ . The asymptotic conditional PMF of the vote-averaging RP-LDA ensemble discriminant (times  $M$ ) according to Moody's model is then

$$p_i(k) = \begin{cases} 1 + \sum_{j=1}^M (-1)^j \binom{M}{j} \prod_{i=1}^j \bar{p}_i, & \text{for } k = 0 \\ \binom{M}{k} \sum_{j=0}^{M-k} \left[ (-1)^j \binom{M-k}{j} \prod_{l=1}^{j+k} \bar{p}_i^{(l)} \right], & \text{for } k = 1, \dots, M \\ 0, & \text{otherwise,} \end{cases}$$

where  $\bar{p}_i^{(j)} = 1 - (1 - \bar{p}_i)(1 - \bar{\rho}_i)^{j-1}$ ,  $j = 2, \dots, M$ . Based on this, the asymptotic probability of misclassification of the uniformly-weighted vote-averaging RP-LDA ensemble discriminant with a threshold of 0.5 is

$$\pi_0 \sum_{k > M/2} p_0(k) + \pi_1 \sum_{k \leq M/2} p_1(k).$$

## A.4 Proof of asymptotic distributions and optimal ensemble construction

### A.4.1 Asymptotic joint distribution of $M$ RP-LDA discriminants

In this section, we prove the asymptotic joint distribution of  $M$  single RP-LDA discriminants as stated in Theorem 3.4.5.

Recall that  $\mathbf{W} = [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_1), \dots, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_M)]^T$  and let

$$\hat{\Sigma}_{\mathbf{R}_k}^{-1} := \mathbf{R}_k^T (\mathbf{R}_k \hat{\Sigma} \mathbf{R}_k^T)^{-1} \mathbf{R}_k, \quad k = 1, \dots, M.$$

Conditioned on  $\{\mathbf{R}_k\}_{k=1}^M$  and the training set  $\mathcal{T}$  and for  $\mathbf{x} \in \mathcal{C}_i$ ,  $\mathbf{W}$  is a Gaussian

vector (through  $\mathbf{x}$ ) with

$$\begin{aligned} \zeta_i &:= \mathbb{E} \left[ \mathbf{W} \mid \{\mathbf{R}_k\}_{k=1}^M, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i \right] \\ &= \begin{bmatrix} \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_1}^{-1} \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \\ \vdots \\ \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_M}^{-1} \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \end{bmatrix}, \quad i = 0, 1, \end{aligned}$$

and covariance  $\boldsymbol{\Pi}$  with entries

$$\text{Var} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k) \mid \mathbf{R}_k, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i] = \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1} \hat{\boldsymbol{\mu}}, \quad k = 1, \dots, M,$$

along the diagonal, and

$$\begin{aligned} &\text{Cov} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k), W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_j) \mid \mathbf{R}_k, \mathbf{R}_j, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i] \\ &= \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_j}^{-1} \hat{\boldsymbol{\mu}}, \quad j, k = 1, \dots, M, \quad j \neq k, \end{aligned}$$

off the diagonal.

From the derivations in Appendix A.1 and Appendix A.2, we know that

$$\begin{aligned} \zeta_i &\asymp \bar{\zeta}_i \\ &= \bar{m}_i \mathbf{1}_M, \quad i = 0, 1, \end{aligned}$$

$$\text{Var} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k) \mid \mathbf{R}_k, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i] \asymp \bar{\sigma}^2(1), \quad \forall k$$

and

$$\text{Cov} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k), W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_j) \mid \mathbf{R}_k, \mathbf{R}_j, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i] \asymp \bar{\sigma}_{M=\infty}^2, \quad \forall j \neq k.$$

Since  $M$  is fixed, then  $\mathbf{\Pi}$  defined above converges pointwise, and so we also have

$$\begin{aligned} \mathbf{\Pi} &\asymp \bar{\mathbf{\Pi}} \\ &= \begin{bmatrix} \bar{\sigma}^2(1) & \bar{\sigma}_{M=\infty}^2 & \cdots & \bar{\sigma}_{M=\infty}^2 \\ \bar{\sigma}_{M=\infty}^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \bar{\sigma}_{M=\infty}^2 \\ \bar{\sigma}_{M=\infty}^2 & \cdots & \bar{\sigma}_{M=\infty}^2 & \bar{\sigma}^2(1) \end{bmatrix} \\ &= (\bar{\sigma}^2(1) - \bar{\sigma}_{M=\infty}^2) \mathbf{I}_M + \bar{\sigma}_{M=\infty}^2 \mathbf{1}_M \mathbf{1}_M^T \end{aligned}$$

Now we prove that  $\mathbf{W}$  converges in distribution to a Gaussian random vector through its characteristic function.

Denote the characteristic function of  $\mathbf{W}$  given  $\mathbf{x} \in \mathcal{C}_i$  by  $\phi_{\mathbf{W},i}(\boldsymbol{\omega})$ . Then

$$\begin{aligned} \phi_{\mathbf{W},i}(\boldsymbol{\omega}) &= \mathbb{E} [\exp(j\boldsymbol{\omega}^T \mathbf{W}) | \mathbf{x} \in \mathcal{C}_i] \\ &= \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M, \mathcal{T}} \left[ \mathbb{E} \left[ \exp(j\boldsymbol{\omega}^T \mathbf{W}) \middle| \{\mathbf{R}_k\}_{k=1}^M, \mathcal{T}, \mathbf{x} \in \mathcal{C}_i \right] \right] \\ &= \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M, \mathcal{T}} \left[ \exp \left( j\boldsymbol{\zeta}_i^T \boldsymbol{\omega} - \frac{1}{2} \boldsymbol{\omega}^T \mathbf{\Pi} \boldsymbol{\omega} \right) \right] \\ &= \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M, \mathcal{T}} \left[ \exp(j(\boldsymbol{\zeta}_i - \bar{\boldsymbol{\zeta}}_i)^T \boldsymbol{\omega}) \exp \left( -\frac{1}{2} \boldsymbol{\omega}^T (\mathbf{\Pi} - \bar{\mathbf{\Pi}}) \boldsymbol{\omega} \right) \exp \left( j\bar{\boldsymbol{\zeta}}_i^T \boldsymbol{\omega} - \frac{1}{2} \boldsymbol{\omega}^T \bar{\mathbf{\Pi}} \boldsymbol{\omega} \right) \right] \\ &\asymp \mathbb{E}_{\{\mathbf{R}_k\}_{k=1}^M, \mathcal{T}} \left[ \exp \left( j\bar{\boldsymbol{\zeta}}_i^T \boldsymbol{\omega} - \frac{1}{2} \boldsymbol{\omega}^T \bar{\mathbf{\Pi}} \boldsymbol{\omega} \right) \right] \\ &= \exp \left( j\bar{\boldsymbol{\zeta}}_i^T \boldsymbol{\omega} - \frac{1}{2} \boldsymbol{\omega}^T \bar{\mathbf{\Pi}} \boldsymbol{\omega} \right) \end{aligned}$$

where the third line follows from the fact that, conditioned on the projections and training, the discriminants are jointly Gaussian, and the second-to-last line is justified through the dominated convergence theorem by the fact that characteristic functions are bounded. The final line reveals a Gaussian characteristic function with mean  $\bar{\boldsymbol{\zeta}}_i$  and covariance  $\bar{\mathbf{\Pi}}$ , thus the vector  $\mathbf{W}$  given  $\mathbf{x} \in \mathcal{C}_i$  is asymptotically Gaussian.

### A.4.2 Asymptotic distribution of the discriminant-averaging RP-LDA finite ensemble discriminant

The asymptotic distribution of the single RP-LDA discriminant follows trivially from the proof of the joint asymptotic distribution of  $M$  RP-LDA discriminants in Appendix A.4.1, by setting  $M = 1$ .

For the general case, using the fact that

$$W_{\text{disc-avg}}(\mathbf{x}, \{\mathbf{R}_k\}_{k=1}^M) \Big|_{\mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M} = \frac{1}{M} \sum_{k=1}^M W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k) \Big|_{\mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M}$$

is Gaussian with mean

$$\frac{1}{M} \sum_{k=1}^M \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_{\mathbf{R}_k}^{-1} \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \asymp \bar{m}_i$$

and variance

$$\begin{aligned} & \frac{1}{M^2} \sum_{k=1}^M \text{Var} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}_k] \\ & + \frac{1}{M^2} \sum_{k \neq j}^M \text{Cov} [W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k), W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_j) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T}, \mathbf{R}_k, \mathbf{R}_j] \\ & \asymp \frac{1}{M} \bar{\sigma}^2(1) + \left(1 - \frac{1}{M}\right) \bar{\sigma}_{M=\infty}^2, \end{aligned}$$

the asymptotic distribution can be proven by convergence of the relevant characteristic function as in Appendix A.4.1.

### A.4.3 Asymptotic distribution of the discriminant-averaging RP-LDA infinite ensemble discriminant

Using the fact that

$$W_{M=\infty}(\mathbf{x}) | \mathbf{x} \in \mathcal{C}_i, \mathcal{T} = \hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} \right] \left( \mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \Big|_{\mathbf{x} \in \mathcal{C}_i, \mathcal{T}}$$

is Gaussian with mean

$$\hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} \right] \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \asymp \bar{m}_i$$

and variance

$$\hat{\boldsymbol{\mu}}^T \mathbb{E}_{\mathbf{R}} \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} \right] \boldsymbol{\Sigma} \mathbb{E}_{\mathbf{R}} \left[ \hat{\boldsymbol{\Sigma}}_{\mathbf{R}}^{-1} \right] \hat{\boldsymbol{\mu}} \asymp \bar{\sigma}_{M=\infty}^2,$$

the asymptotic distribution can be proven by convergence of the relevant characteristic function as in Appendix A.4.1.

#### A.4.4 Asymptotic distribution of the vote-averaging RP-LDA ensemble discriminant

The class-conditional discriminant

$$MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i = \sum_{k=1}^M \mathbb{1} \{W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k)\} | \mathbf{x} \in \mathcal{C}_i$$

is clearly a sum of correlated Bernoullis. The probability of success for each Bernoulli and the correlations between Bernoullis vary through their random projections. Thus,  $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i$  is a correlated Binomial random variable with varying probability of success for each trial and varying correlations between trials.

The PMF of  $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i$ , given by

$$P[MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i = m], \quad m = 0, 1, \dots, M,$$

can be obtained exactly as a function of the underlying discriminants

$$\{W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_k) | \mathbf{x} \in \mathcal{C}_i\}_{k=1}^M$$

by summing over all probabilities where exactly  $m$  single RP-LDA discriminants

are greater than zero. For example,

$$\begin{aligned}
& P[MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i = 1] = \\
& P[W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_1) > 0, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_2) < 0, \dots, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_M) < 0] \\
& + P[W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_1) < 0, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_2) > 0, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_3) < 0, \dots] + \dots \\
& + P[W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_1) < 0, \dots, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_{M-1}) < 0, W_{\text{RP-LDA}}(\mathbf{x}, \mathbf{R}_M) > 0].
\end{aligned} \tag{A.31}$$

Moreover, the corresponding CDF is simply a cumulative sum of the PMF.

We have from Theorem 3.4.5 that the class-conditional joint distribution of  $M$  single RP-LDA discriminants converges to a Gaussian with mean  $\bar{\boldsymbol{\zeta}}_i$  and covariance  $\bar{\boldsymbol{\Pi}}$ . Thus the PMF, and, as a result, the CDF of  $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i$ , can be computed asymptotically based on the limiting distribution. Formally, let  $\bar{\mathbf{W}}_i = [\bar{W}_{1,i}, \dots, \bar{W}_{M,i}]^T$  denote a Gaussian with  $\bar{\boldsymbol{\zeta}}_i$  and covariance  $\bar{\boldsymbol{\Pi}}$ . Since

$$\lim_{n,p,d \rightarrow \infty} P[MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i = m] - P\left[\sum_{k=1}^M \mathbb{1}\{\bar{W}_{k,i}\} = m\right] = 0$$

through the fact that the left-hand side can be expressed as the limit on a sum of probabilities involving single RP-LDA discriminants (as in (A.31)) and also the underlying convergence in distribution of these discriminants shown in Theorem 3.4.5, then  $\forall x \in \mathbb{R}$ ,

$$\lim_{n,p,d \rightarrow \infty} P[MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i \leq x] - P\left[\sum_{k=1}^M \mathbb{1}\{\bar{W}_{k,i}\} \leq x\right] = 0, \tag{A.32}$$

which is convergence in distribution.

The term  $\sum_{k=1}^M \mathbb{1}\{\bar{W}_{k,i}\}$  is a correlated Binomial consisting of  $M$  trials. It is straightforward to compute the probability of success of its trials and correlations

between its trials as a function of the distribution of  $\bar{W}_i$ . Because of the structure of  $\bar{\zeta}_i$  and  $\bar{\Pi}$ , the probabilities of success and correlations are constants denoted  $\bar{p}_i$  and  $\bar{\rho}_i$ .

It is easy to show that  $\bar{\rho}_i$  is always positive, which is reasonable, as more than two variables cannot be simultaneously negatively correlated.

Note that (A.32) gives a way to approximate the PMF of  $MW_{\text{vote-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) | \mathbf{x} \in \mathcal{C}_i$ . Since  $\{\bar{W}_{k,i}\}_{k=1}^M$  are identically distributed, computing the asymptotic PMF becomes a counting problem. These computations, however, still involve numerical integration, and this can become restrictive when  $M$  is large, and for that reason we propose the approximation of the asymptotic PMF by Moody's correlated Binomial PMF in Appendix A.3.3.

## A.5 Derivation of G-estimators

This section derives the G-estimators of the most common metrics of binary classification. These rely on building blocks  $\hat{m}_i$ ,  $i = 0, 1$ ,  $\hat{\sigma}^2(1)$ , and  $\hat{\sigma}_{M=\infty}^2$ . As  $\hat{m}_i$ ,  $i = 0, 1$  and  $\hat{\sigma}_{M=\infty}^2$  were derived in detail in our previous work (Niyazi et al., 2020b), we consider only  $\hat{\sigma}^2(1)$  in the current work. Section A.5.1 derives  $\hat{\sigma}^2(1)$ , while Section A.5.2 proves Theorem 3.5.2. Additionally, Section A.5.3 derives the approximation of the infinite to finite error ratio used to solve for  $M$  in the heuristic introduced in Section 3.5.2.

### A.5.1 Derivation of $\hat{\sigma}^2(1)$

The first step is to derive the quantity  $\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)$  as a function of the training (as opposed to true statistics as was done in Appendix A.1). From the system of equations (A.14), we have

$$1 - \frac{p}{d} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)} \frac{1}{d} \text{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)} \mathbf{I}_p \right)^{-1} \right\} = 0. \quad (\text{A.33})$$

The trace term on the left-hand side can be rewritten as

$$\begin{aligned} \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)} \frac{1}{d} \text{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)} \mathbf{I}_p \right)^{-1} \right\} &= \frac{1}{d} \text{tr} \left\{ \left( \lim_{\beta \rightarrow 0} \tilde{\nu}(\beta) \mathbf{D} + \mathbf{I}_p \right)^{-1} \right\} \\ &= \frac{1}{d} \sum_{i=1}^p \frac{1}{1 + \lim_{\beta \rightarrow 0} \tilde{\nu}(\beta) \lambda_i(\hat{\Sigma})}. \end{aligned}$$

Now consider the monotonically decreasing function

$$f(x) = 1 - \frac{p}{d} + \frac{1}{d} \sum_{i=1}^p \frac{1}{1 + x \lambda_i(\hat{\Sigma})}.$$

As  $x \rightarrow 0$ ,  $f(x) \rightarrow 1$  and as  $x \rightarrow \infty$ ,  $f(x) \rightarrow 1 - \frac{p}{d} < 0$ , when  $p > d$ , which is the typical use-case. Therefore,  $f(x)$  has a unique root  $x^*$  and  $\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta) = x^*$ . Since (A.33) can be rewritten as

$$1 - \frac{1}{d} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)} \mathbf{I}_p \right)^{-1} \right\} = 0,$$

then overall, the G-estimator of  $\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)$ , denoted  $\hat{\nu}$ , is such that

$$1 - \frac{1}{d} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\} = 0. \quad (\text{A.34})$$

Now taking the limit as  $\beta$  goes to zero on the intermediate convergence in (A.22) and replace  $\lim_{\beta \rightarrow 0} \tilde{\nu}(\beta)$  by its G-estimator  $\hat{\nu}$ , we have

$$\begin{aligned} \sigma^2(1) &\asymp \hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \boldsymbol{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}} \\ &\quad + \frac{1}{\hat{\nu}^2} \frac{\frac{1}{d} \text{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \boldsymbol{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}}{1 - \frac{1}{d} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-2} \hat{\boldsymbol{\mu}}. \end{aligned}$$

Only two terms involve the true statistic  $\boldsymbol{\Sigma}$ , while the remaining terms are func-

tions of the sample statistics. These two terms can be estimated as

$$\begin{aligned} & \left( \frac{1}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \right)^2 \hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}} \\ & \asymp \hat{\boldsymbol{\mu}}^T \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \Sigma \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\boldsymbol{\mu}} \end{aligned}$$

and

$$\begin{aligned} & \left( \frac{1}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}} \right)^2 \frac{1}{p} \text{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\} \\ & \asymp \frac{1}{p} \text{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \Sigma \left( \hat{\Sigma} + \frac{1}{\hat{\nu}} \mathbf{I}_p \right)^{-1} \right\}. \end{aligned}$$

The proof uses the same techniques used in Section B of Appendix B of (Niyazi et al., 2020a). The same growth regime assumptions stated at the beginning of Section 3.4 apply here.

### A.5.2 Proof of Theorem 3.5.2

First we derive the exact probabilities as follows:

$$\begin{aligned} TPR &= P[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) > 0 | \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M, \mathbf{x} \in \mathcal{C}_1] \\ &= \Phi \left( \frac{m_1}{\sqrt{\sigma^2(M)}} \right), \end{aligned}$$

$$\begin{aligned} TNR &= P[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) < 0 | \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M, \mathbf{x} \in \mathcal{C}_0] \\ &= \Phi \left( -\frac{m_0}{\sqrt{\sigma^2(M)}} \right), \end{aligned}$$

$$\begin{aligned} FPR &= P[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) > 0 | \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M, \mathbf{x} \in \mathcal{C}_0] \\ &= \Phi\left(\frac{m_0}{\sqrt{\sigma^2(M)}}\right), \end{aligned}$$

and

$$\begin{aligned} FNR &= P[W_{\text{disc-avg}}(\mathbf{x}, \mathbf{R}_1, \dots, \mathbf{R}_M) < 0 | \mathcal{T}, \{\mathbf{R}_k\}_{k=1}^M, \mathbf{x} \in \mathcal{C}_1] \\ &= \Phi\left(-\frac{m_1}{\sqrt{\sigma^2(M)}}\right). \end{aligned}$$

We are then able to substitute the G-estimators for each of the quantities  $m_0$ ,  $m_1$ , and  $\sigma^2(M)$  in the above expressions by a similar argument to that presented for Lemma 2 in (Niyazi et al., 2020b). The G-estimators for the following quantities are derived in a similar fashion using the G-estimators for the above quantities:

$$\varepsilon = \pi_0 \text{FPR} + \pi_1 \text{FNR},$$

$$\text{PPV} = \frac{\pi_1 \text{TPR}}{\pi_0 \text{FPR} + \pi_1 \text{TPR}},$$

and

$$\text{NPV} = \frac{\pi_0 \text{TNR}}{\pi_0 \text{TNR} + \pi_1 \text{FNR}}.$$

### A.5.3 Derivation of the heuristic approximation

Let  $\hat{\varepsilon}_{M=\infty}$  and  $\hat{\varepsilon}(M)$  denote the G-estimators of the probability of misclassification of the infinite and finite discriminant-averaging ensembles, respectively, where the

latter consists of  $M$  randomly-projected LDA discriminants. Then,

$$\frac{\hat{\varepsilon}_{M=\infty}}{\hat{\varepsilon}(M)} = \frac{\hat{\pi}_0 \Phi\left(\frac{\hat{m}_0}{\sqrt{\hat{\sigma}_{M=\infty}^2}}\right) + \hat{\pi}_1 \Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}_{M=\infty}^2}}\right)}{\hat{\pi}_0 \Phi\left(\frac{\hat{m}_0}{\sqrt{\hat{\sigma}^2(M)}}\right) + \hat{\pi}_1 \Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)}.$$

By assuming equal priors,  $\pi_0 = \pi_1$ ,  $n_0 = n_1$ ,  $\hat{\pi}_0 = \hat{\pi}_1$ , and  $\hat{m}_0 = -\hat{m}_1$ . Then

$$\begin{aligned} \frac{\hat{\varepsilon}_{M=\infty}}{\hat{\varepsilon}(M)} &= \frac{\Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}_{M=\infty}^2}}\right)}{\Phi\left(-\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)} \\ &= \frac{Q\left(\frac{\hat{m}_1}{\sqrt{\hat{\sigma}_{M=\infty}^2}}\right)}{Q\left(\frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}\right)}, \end{aligned} \tag{A.35}$$

where  $Q(\cdot)$  is the complementary CDF of a standard Gaussian random variable.

The approximation

$$Q(x) \approx \frac{1/\sqrt{2\pi} \exp(-x^2/2)}{x}, \quad x > 0,$$

follows from the using the right-hand side of the inequality

$$\frac{x}{1+x^2} 1/\sqrt{2\pi} \exp(-x^2/2) < Q(x) < \frac{1/\sqrt{2\pi} \exp(-x^2/2)}{x}, \quad x > 0$$

which becomes tighter with increasing  $x$  (Borjesson and Sundberg, 1979). Ap-

plying this inequality to (A.35) with  $x := \frac{\hat{m}_1}{\sqrt{\hat{\sigma}_{M=\infty}^2}}$  and  $y := \frac{\hat{m}_1}{\sqrt{\hat{\sigma}^2(M)}}$ , we obtain

$$\frac{\hat{\varepsilon}_{M=\infty}}{\hat{\varepsilon}(M)} \approx \frac{1/\sqrt{2\pi} \exp(-x^2/2) / x}{1/\sqrt{2\pi} \exp(-y^2/2) / y}.$$

Setting this to  $\psi$  and solving for  $y$  (which is a function of the desired  $M$ ), we

have

$$y^{-1} \exp(-y^2/2) = \frac{x^{-1} \exp(-x^2/2)}{\psi}.$$

Squaring and inverting both sides of this equation yields

$$y^2 \exp(y^2) = \psi^2 x^2 \exp(x^2)$$

which can be solved for  $y^2 = \frac{\hat{m}_1^2}{\hat{\sigma}^2(M)}$  by applying the principal branch of the Lambert W function,  $W_0(\cdot)$ , to both sides (since they are positive). Then

$$y^2 = \frac{\hat{m}_1^2}{\hat{\sigma}^2(M)} = W_0(\psi^2 x^2 \exp(x^2)).$$

By making use of the fact that  $\hat{\sigma}^2(M) = \frac{1}{M}\hat{\sigma}^2(1) + (1 - \frac{1}{M})\hat{\sigma}_{M=\infty}^2$  and  $x^2 = \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2}$ , while solving for  $M$ , we have

$$M \approx \text{ceil} \left( \frac{(\hat{\sigma}^2(1) - \hat{\sigma}_{M=\infty}^2) W_0 \left( \psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \exp \left( \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \right) \right)}{\hat{m}_1^2 - \hat{\sigma}_{M=\infty}^2 W_0 \left( \psi^2 \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \exp \left( \frac{\hat{m}_1^2}{\hat{\sigma}_{M=\infty}^2} \right) \right)} \right).$$

## Appendix B

### Proofs for Chapter 4

#### B.1 Main result proofs

##### B.1.1 Proof of Theorem 1

The G-estimator of (4.9) is obtained by substituting G-estimators of the numerator and denominator into the expression (4.9). We start with deriving the G-estimator of the numerator term,  $\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$ ,  $i = 0, 1$ , followed by the G-estimator of the denominator term,  $\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}}$ .

## Numerator

To estimate the numerators in (4.9), notice that  $\tilde{\Sigma}_{\text{WS}}$  and  $\hat{\boldsymbol{\mu}}_i$ ,  $i = 0, 1$ , are independent since  $\hat{\Sigma}$  and  $\hat{\boldsymbol{\mu}}_i$ ,  $i = 0, 1$ , are independent and  $\tilde{\Sigma}_{\text{WS}}$  is a function of  $\hat{\Sigma}$ . This allows us to take the limits with respect to  $\hat{\boldsymbol{\mu}}_i$ ,  $i = 0, 1$ , independently of  $\tilde{\Sigma}_{\text{WS}}$ . First, express the numerator term as

$$\hat{\boldsymbol{\mu}}^T \tilde{\Sigma}_{\text{WS}} \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) = \hat{\boldsymbol{\mu}}^T \tilde{\Sigma}_{\text{WS}} \left( \hat{\boldsymbol{\mu}}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \hat{\boldsymbol{\mu}}^T \tilde{\Sigma}_{\text{WS}} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i), \quad i = 0, 1.$$

In this form, we need to estimate  $\hat{\boldsymbol{\mu}}^T \tilde{\Sigma}_{\text{WS}} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i)$ ,  $i = 0, 1$ , since it involves the true means which are unknown in practice. Using the fact that  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i - 1}$  for some  $\mathbf{Z} \in \mathbb{R}^{p \times (n_i - 1)}$  with i.i.d. Gaussian entries and  $\frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i - 1} \sim \mathcal{N} \left( \mathbf{0}_p, \frac{\boldsymbol{\Sigma}}{n_i - 1} \right)$ , we have

$$\hat{\boldsymbol{\mu}}^T \tilde{\Sigma}_{\text{WS}} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i) \asymp (-1)^i \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \tilde{\Sigma}_{\text{WS}} \right\}, \quad i = 0, 1. \quad (\text{B.1})$$

Leveraging this convergence result allows us to estimate the right-hand side term instead. This is done using the same approach as Rubio and Mestre (2009) which is detailed in what follows.

The term  $\frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \tilde{\Sigma}_{\text{WS}} \right\}$  is expressed in terms of a contour integral over a contour  $\mathcal{C}$  which encloses the eigenvalues,  $\{l_i\}_{i=1}^p$  (more specifically  $\mathcal{C}$  is an open subset of  $\mathbb{C}$  containing the interval  $[-\epsilon, \infty)$ ,  $\epsilon > 0$ ), using the Cauchy integral formula

$$f(a) = \frac{1}{2\pi i} \oint_{\mathcal{A}} \frac{f(z)}{z - a} dz, \quad (\text{B.2})$$

where  $f$  is analytic in  $\mathcal{A}$  and  $a \in \mathcal{A}$ . To do this, first express

$$\begin{aligned} \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \tilde{\Sigma}_{\text{WS}} \right\} &= \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \left( \sum_{j=1}^p d_j^{\text{WS}} \mathbf{u}_j \mathbf{u}_j^T \right) \right\} \\ &= \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \left( \sum_{j=1}^p \boldsymbol{\alpha}^T \mathbf{h}(l_j) \mathbf{u}_j \mathbf{u}_j^T \right) \right\}, \end{aligned} \quad (\text{B.3})$$

where  $\mathbf{u}_j$  is the  $j^{\text{th}}$  column of  $\mathbf{U}$ . Now applying (B.2) to  $\boldsymbol{\alpha}^T \mathbf{h}(l_j)$  in (B.3), we have

$$\begin{aligned} \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \right\} &= \frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \left( \sum_{j=1}^p \mathbf{u}_j \mathbf{u}_j^T \frac{1}{z - l_j} \right) \right\} \boldsymbol{\alpha}^T \mathbf{h}(z) dz \\ &= -\frac{1}{2\pi i} \oint_{\mathcal{C}} \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \mathbf{Q}(z) \right\} \boldsymbol{\alpha}^T \mathbf{h}(z) dz \end{aligned} \quad (\text{B.4})$$

where  $\mathbf{Q}(z) = \left( \hat{\boldsymbol{\Sigma}} - z \mathbf{I}_p \right)^{-1}$ . Notice that this step has introduced the resolvent into our expression (which we know how to deal with). Also note that the Cauchy integral theorem requires that  $\boldsymbol{\alpha}^T \mathbf{h}(z)$  be analytic in  $\mathcal{C}$ , hence the inclusion of assumption (f) at the beginning of Section 4.4.2.

We can estimate the integrand in (B.4) as (see Appendix B in Niyazi et al. (2020a))

$$\boldsymbol{\alpha}^T \mathbf{h}(z) \frac{\frac{1}{n_i - 1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z) \right\}}{1 - \frac{1}{n - 2} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z) \right\}} \asymp \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \mathbf{Q}(z) \right\} \boldsymbol{\alpha}^T \mathbf{h}(z).$$

This step requires that  $\boldsymbol{\alpha}^T \mathbf{h}(z)$ . This is guaranteed by the assumptions (f) and ((g)) stated at the beginning of Section 4.4.2.

From this it follows that the estimator is

$$-\frac{1}{2\pi i} \oint_{\mathcal{C}} \boldsymbol{\alpha}^T \mathbf{h}(z) \frac{\frac{1}{n_i - 1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z) \right\}}{1 - \frac{1}{n - 2} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z) \right\}} dz \asymp \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \right\} \quad (\text{B.5})$$

We now need to evaluate the contour integral. To do this, we must find the poles of the integrand of the left-hand side of (B.5), determine their orders, and finally compute the residues. We consider each of the two cases when  $\rho = p$  and  $\rho < p$  separately.

- $\rho = p$

First note that  $h_k$  is analytic by assumption (f) and therefore contributes no singularities to the integrand. We can rewrite the remainder of the

integrand as:

$$\frac{\frac{1}{n_i-1} \text{tr} \left\{ \hat{\Sigma} \mathbf{Q}(z) \right\}}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\Sigma} \mathbf{Q}(z) \right\}} = \frac{\frac{1}{n_i-1} \sum_{j=1}^p l_j \prod_{\substack{m=1 \\ m \neq j}}^p (l_m - z)}{\prod_{j=1}^p (l_j - z) - \frac{1}{n-2} \sum_{j=1}^p l_j \prod_{\substack{m=1 \\ m \neq j}}^p (l_m - z)} \quad (\text{B.6})$$

In this form, the poles are clearly the  $p$  solutions to the equation

$$\prod_{j=1}^p (l_j - z) = \frac{1}{n-2} \sum_{j=1}^p l_j \prod_{\substack{m=1 \\ m \neq j}}^p (l_m - z),$$

or equivalently

$$\frac{1}{n-2} \sum_{j=1}^p \frac{l_j}{l_j - z} = 1. \quad (\text{B.7})$$

According to Lemma 8.1 of Couillet and Debbah (2011), the zeros,  $\{\nu_i\}_{i=1}^p$ , of (B.7) (which are the poles of (B.6)) are the eigenvalues of  $\mathbf{L} - \left(\sqrt{\frac{1}{n-2}}\right) \left(\sqrt{\frac{1}{n-2}}\right)^T$ , where  $\mathbf{l} = [l_1, \dots, l_p]^T$ . Note that  $\{\nu_i\}_{i=1}^p$  satisfy  $\nu_1 < l_1 < \nu_2 < \dots < \nu_p < l_p$  Rubio and Mestre (2009) (where the poles and sample eigenvalues are sorted in ascending order), which means that they are enclosed by  $\mathcal{C}$  and thus contribute to the residue. This also means that the poles are simple since they are distinct. Next, we compute residues.

We evaluate the residues of the integrand at the poles we just derived and sum them up. The Residue Theorem states that for a complex function  $f(z)$  which is analytic on an open subset, except for singularities  $z_1, \dots, z_n$ , the contour integral over a curve  $\mathcal{C}$  enclosing these singular points is only due to the contribution of these singular points, i.e.,

$$\oint_{\mathcal{C}} f(z) dz = 2\pi i \sum_{j=0}^n \text{Res}(f, z_j),$$

where  $\text{Res}(f, z_j)$  is the residue of the Laurent series expansion of  $f(z)$  about the singularity  $z_j$ , i.e., it is the coefficient  $b_1$  of  $\frac{1}{z-z_j}$  in the Laurent series expansion of  $f(z)$  about  $z_j$ . For simple poles, we have for a function of the

form  $f(z) = \frac{g(z)}{h(z)}$ , that the residue of  $f$  at the pole  $z = c$  is given by

$$\text{Res}(f, c) = \frac{g(c)}{h'(c)}.$$

Therefore, when  $\rho = p$ ,

$$\begin{aligned} -\frac{1}{2\pi i} \oint_{\mathcal{C}} \boldsymbol{\alpha}^T \mathbf{h}(z) \frac{\frac{1}{n_i-1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z) \right\}}{1 - \frac{1}{n-2} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z) \right\}} dz &= \sum_{l=1}^p \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(\nu_l) \right\}}{\frac{1}{n-2} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}^2(\nu_l) \right\}} \\ &\asymp \frac{1}{n_i - 1} \text{tr} \left\{ \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \right\}. \end{aligned} \quad (\text{B.8})$$

So overall, using (B.1) and (B.8), we have

$$(-1)^i \sum_{l=1}^p \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(\nu_l) \right\}}{\frac{1}{n-2} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}^2(\nu_l) \right\}} \asymp \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i), \quad i = 0, 1,$$

and

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \hat{\boldsymbol{\mu}}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + (-1)^i \sum_{l=1}^p \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(\nu_l) \right\}}{\frac{1}{n-2} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}^2(\nu_l) \right\}} \\ \asymp \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right), \quad i = 0, \end{aligned} \quad (\text{B.9})$$

To unify the notation between the case when  $\hat{\boldsymbol{\Sigma}}$  is full rank and when it is singular, we adopt the generalized definitions (4.10), (4.11), (4.12), (4.13), and (4.14), which recover the quantities  $\mathbf{U}$ ,  $\mathbf{I}$ ,  $\mathbf{L}$ ,  $\hat{\boldsymbol{\Sigma}}$ , and  $\mathbf{z}$ , respectively, when  $\rho := \text{rank} \left\{ \hat{\boldsymbol{\Sigma}} \right\} = p$ , i.e., when  $\hat{\boldsymbol{\Sigma}}$  is full rank. Thus, the left-hand side of (B.9) can be equivalently written in terms of (4.10)-(4.14). This is important for the presentation of the final result in Theorem 4.4.1.

- $\rho < p$

When  $\rho < p$ , we can further simplify the right-hand side of (B.6) by making use of the fact that some of the sample eigenvalues are zero. Under our

Gaussian data assumptions there are exactly  $n - 2$  non-zero eigenvalues and  $p - n + 2$  zero eigenvalues. Let  $\{l_i\}_{i=1}^{p-n+2}$  be the zero eigenvalues and  $\{l_i\}_{i=p-n+3}^p$  be the non-zero eigenvalues. Then,

$$\prod_{j=1}^p (l_j - z) = (-z)^{p-n+2} \prod_{j=p-n+3}^p (l_j - z)$$

and

$$\sum_{j=1}^p l_j \prod_{\substack{m=1 \\ m \neq j}}^p (l_m - z) = (-z)^{p-n+2} \sum_{j=p-n+3}^p l_j \prod_{\substack{m=p-n+3 \\ m \neq j}}^p (l_m - z).$$

Therefore, when  $p > n$ , the factor of  $(-z)^{p-n+2}$  in the numerator and denominator of (B.6) cancels out, and the expression further simplifies to

$$\frac{\frac{1}{n-1} \sum_{j=p-n+3}^p l_j \prod_{\substack{m=p-n+3 \\ m \neq j}}^p (l_m - z)}{\prod_{j=p-n+3}^p (l_j - z) - \frac{1}{n-2} \sum_{j=p-n+3}^p l_j \prod_{\substack{m=p-n+3 \\ m \neq j}}^p (l_m - z)}. \quad (\text{B.10})$$

The poles are thus the  $n - 2$  solutions to the equation

$$\prod_{j=p-n+3}^p (l_j - z) = \frac{1}{n-2} \sum_{j=p-n+3}^p l_j \prod_{\substack{m=p-n+3 \\ m \neq j}}^p (l_m - z).$$

which we denote by  $\{\nu_i\}_{i=1}^{n-2}$ . The solutions to (B.10) are equivalent to the solutions of

$$\frac{1}{n-2} \sum_{i=p-n+3}^p \frac{l_i}{l_i - z} = 1. \quad (\text{B.11})$$

Note that we can immediately see from (B.11) that  $z = 0$  is a pole. This was not the case when  $\rho = p$ . So we have  $n - 2$  poles,  $\{\nu_i\}_{i=1}^{n-2}$ , which satisfy the inequality  $0 = \nu_1 = l_1 = \dots = l_{p-n+2} < \nu_2 < l_{p-n+3} < \dots < \nu_{n-2} < l_p$  (sorting the poles and eigenvalues in ascending order). We can compute the poles using Lemma 8.1 in Couillet and Debbah (2011). Note it is very important to use only the non-zero sample eigenvalues in the computation so we redefine  $\mathbf{L}$  and  $\mathbf{l}$  as being composed of only the non-zero eigenvalues when  $\rho < p$  as in (4.11) and (4.12) and we apply Lemma 8.1 to

$\bar{\mathbf{L}} - \left(\sqrt{\frac{\bar{\mathbf{1}}}{n-2}}\right) \left(\sqrt{\frac{\bar{\mathbf{1}}}{n-2}}\right)^T$ . The poles are simple and are all enclosed by the contour  $\mathcal{C}$ . The integral is now

$$\frac{1}{2\pi i} \oint_{\mathcal{C}} \boldsymbol{\alpha}^T \mathbf{h}(z) \frac{\frac{1}{n_i-1} \sum_{j=p-n+3}^p l_j \prod_{\substack{m=p-n+3 \\ m \neq j}}^p (l_m - z)}{\prod_{j=p-n+3}^p (l_j - z) - \frac{1}{n-2} \sum_{j=p-n+3}^p l_j \prod_{\substack{m=p-n+3 \\ m \neq j}}^p (l_m - z)} dz. \quad (\text{B.12})$$

We want to write this in terms of matrices, so that the final estimator is in a nice convenient form. We do this using the quantities defined in (4.10)-(4.14). Dividing the numerator and denominator of the integrand in (B.12) by  $\prod_{i=p-n+3}^p (l_i - z_1)$ , we have

$$\frac{1}{2\pi i} \oint_{\mathcal{C}} \boldsymbol{\alpha}^T \mathbf{h}(z) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(z) \}}{1 - \frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(z) \}} dz. \quad (\text{B.13})$$

Now, computing the residues based on (B.13), we have, for the case when  $\rho < p$ ,

$$\begin{aligned} -\frac{1}{2\pi i} \oint_{\mathcal{C}} \boldsymbol{\alpha}^T \mathbf{h}(z) \frac{\frac{1}{n_i-1} \text{tr} \{ \hat{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(z) \}}{1 - \frac{1}{n-2} \text{tr} \{ \hat{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(z) \}} dz &= \sum_{l=1}^{n-2} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_l) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \}} \\ &\asymp \frac{1}{n_i-1} \text{tr} \{ \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \}, \end{aligned} \quad (\text{B.14})$$

where  $\{\nu_i\}_{i=1}^{n-2}$  are the eigenvalues of  $\bar{\mathbf{L}} - \left(\sqrt{\frac{\bar{\mathbf{1}}}{n-2}}\right) \left(\sqrt{\frac{\bar{\mathbf{1}}}{n-2}}\right)^T$ .

So overall, using (B.1) and (B.14), we have

$$(-1)^i \sum_{l=1}^{n-2} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_l) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \}} \asymp \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i), \quad i = 0, 1,$$

and

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \hat{\boldsymbol{\mu}}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + (-1)^i \sum_{l=1}^{n-2} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_l) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \}} \\ \asymp \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \boldsymbol{\mu}_i - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right), \quad i = 0, 1. \end{aligned}$$

Combining the two cases when  $\hat{\boldsymbol{\Sigma}}$  is full rank and singular together, the unified G-estimator for the numerator term is

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \hat{\boldsymbol{\mu}}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + (-1)^i \sum_{l=1}^{\rho} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_l) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \}} \\ \asymp \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \boldsymbol{\mu}_i - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right), \quad i = 0, 1. \quad (\text{B.15}) \end{aligned}$$

Note that this expression also generalizes the result when  $\rho < p$  to non-Gaussian data where there are  $\rho \neq n - 2$  non-zero eigenvalues in general .

## Denominator

Now we derive the G-estimator of the denominator term,  $\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}}$ , of (4.9). Using the Cauchy integral formula as before, we have

$$\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}} = \frac{1}{(2\pi i)^2} \oint_{\mathcal{C}} \oint_{\mathcal{C}} \hat{\boldsymbol{\mu}}^T \mathbf{Q}(z_1) \boldsymbol{\Sigma} \mathbf{Q}(z_2) \hat{\boldsymbol{\mu}} \boldsymbol{\alpha}^T \mathbf{h}(z_1) \boldsymbol{\alpha}^T \mathbf{h}(z_2) dz_1 dz_2.$$

Let

$$w(z) := 1 - \frac{1}{n-2} \text{tr} \{ \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z) \}.$$

We estimate the integrand using the convergence relation (see Appendix B of Niyazi et al. (2020a))

$$\frac{1}{w(z_1)w(z_2)} \boldsymbol{\mu}^T \mathbf{Q}(z_1) \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z_2) \boldsymbol{\mu} \boldsymbol{\alpha}^T \mathbf{h}(z_1) \boldsymbol{\alpha}^T \mathbf{h}(z_2) \asymp \hat{\boldsymbol{\mu}}^T \mathbf{Q}(z_1) \boldsymbol{\Sigma} \mathbf{Q}(z_2) \hat{\boldsymbol{\mu}} \boldsymbol{\alpha}^T \mathbf{h}(z_1) \boldsymbol{\alpha}^T \mathbf{h}(z_2),$$

and compute

$$\frac{1}{(2\pi i)^2} \oint_{\mathcal{C}} \oint_{\mathcal{C}} \frac{1}{w(z_1)w(z_2)} \boldsymbol{\mu}^T \mathbf{Q}(z_1) \hat{\boldsymbol{\Sigma}} \mathbf{Q}(z_2) \boldsymbol{\mu} \boldsymbol{\alpha}^T \mathbf{h}(z_1) \boldsymbol{\alpha}^T \mathbf{h}(z_2) dz_1 dz_2$$

to get our estimator. Following similar steps as in the derivation for the numerator, we obtain

$$\sum_{l=1}^{\rho} \sum_{k=1}^{\rho} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \boldsymbol{\alpha}^T \mathbf{h}(\nu_k) \frac{\hat{\boldsymbol{\mu}}^T \bar{\mathbf{Q}}(\nu_l) \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_k) \hat{\boldsymbol{\mu}}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \} \frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_k) \}} \asymp \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}}. \quad (\text{B.16})$$

where  $\{\nu_i\}_{i=1}^{\rho}$  are the eigenvalues of  $\bar{\mathbf{L}} - \left( \sqrt{\frac{1}{n-2}} \right) \left( \sqrt{\frac{1}{n-2}} \right)^T$ .

Combining the results in (B.15) and (B.16) yields

$$\sum_{i=0}^1 \hat{\pi}_i \Phi \left( (-1)^i \frac{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \hat{\boldsymbol{\mu}}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + (-1)^i \sum_{l=1}^{\rho} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_l) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \}} + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\sum_{l=1}^{\rho} \sum_{k=1}^{\rho} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \boldsymbol{\alpha}^T \mathbf{h}(\nu_k) \frac{\hat{\boldsymbol{\mu}}^T \bar{\mathbf{Q}}(\nu_l) \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_k) \hat{\boldsymbol{\mu}}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \} \frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_k) \}}}} \right). \quad (\text{B.17})$$

To express (B.17) in a compact form, first we can show that

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \hat{\boldsymbol{\mu}}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) &= (-1)^{i+1} \frac{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}}}{2} \\ &= \frac{(-1)^{i+1}}{2} \boldsymbol{\alpha}^T [\mathbf{h}(l_1) \cdots \mathbf{h}(l_p)] \mathbf{b}, \quad i = 0, 1, \end{aligned}$$

where  $\mathbf{b} = \mathbf{U}^T \hat{\boldsymbol{\mu}} \circ \mathbf{U}^T \hat{\boldsymbol{\mu}}$ . Also, we have

$$\sum_{l=1}^{\rho} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_l) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \}} = \boldsymbol{\alpha}^T [\mathbf{h}(\nu_1) \cdots \mathbf{h}(\nu_{\rho})] \mathbf{c}_i, \quad i = 0, 1,$$

where  $\mathbf{c}_i = \left[ \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_1) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_1) \}}, \dots, \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_{\rho}) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_{\rho}) \}} \right]^T$ . Letting the  $r \times p$  matrix  $\mathbf{H}_1 := [\mathbf{h}(l_1) \cdots \mathbf{h}(l_p)]$  and the  $r \times \rho$  matrix  $\mathbf{H}_{\nu} := [\mathbf{h}(\nu_1) \cdots \mathbf{h}(\nu_{\rho})]$ , we have in the

numerator of (B.17)

$$\begin{aligned}
& \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left( \hat{\boldsymbol{\mu}}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + (-1)^i \sum_{l=1}^{\rho} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \frac{\frac{1}{n_i-1} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_l) \}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \}} \hat{\pi}_0 \\
&= \frac{(-1)^{i+1}}{2} \boldsymbol{\alpha}^T [\mathbf{h}(l_1) \cdots \mathbf{h}(l_p)] \mathbf{b} + (-1)^i \boldsymbol{\alpha}^T [\mathbf{h}(\nu_1) \cdots \mathbf{h}(\nu_\rho)] \mathbf{c}_i \\
&= \frac{(-1)^{i+1}}{2} \boldsymbol{\alpha}^T \mathbf{H}_1 \mathbf{b} + (-1)^i \boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{c}_i \\
&= \boldsymbol{\alpha}^T \left[ \frac{(-1)^{i+1}}{2} \mathbf{H}_1 \mathbf{b} + (-1)^i \mathbf{H}_\nu \mathbf{c}_i \right]. \tag{B.18}
\end{aligned}$$

Similarly, the term in the denominator of (B.17) can be rewritten as

$$\begin{aligned}
& \sum_{l=1}^{\rho} \sum_{k=1}^{\rho} \boldsymbol{\alpha}^T \mathbf{h}(\nu_l) \boldsymbol{\alpha}^T \mathbf{h}(\nu_k) \frac{\hat{\boldsymbol{\mu}}^T \bar{\mathbf{Q}}(\nu_l) \bar{\boldsymbol{\Sigma}}(\mathbf{X}_0, \mathbf{X}_1) \bar{\mathbf{Q}}(\nu_k) \hat{\boldsymbol{\mu}}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}}(\mathbf{X}_0, \mathbf{X}_1) \bar{\mathbf{Q}}^2(\nu_l) \} \frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}}(\mathbf{X}_0, \mathbf{X}_1) \bar{\mathbf{Q}}^2(\nu_k) \}} \\
&= \boldsymbol{\alpha}^T [\mathbf{h}(\nu_1) \cdots \mathbf{h}(\nu_\rho)] \boldsymbol{\Pi} \begin{bmatrix} \mathbf{h}^T(\nu_1) \\ \vdots \\ \mathbf{h}^T(\nu_\rho) \end{bmatrix} \boldsymbol{\alpha} \\
&= \boldsymbol{\alpha}^T \mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}, \tag{B.19}
\end{aligned}$$

where  $[\boldsymbol{\Pi}]_{l,k} = \frac{\hat{\boldsymbol{\mu}}^T \bar{\mathbf{Q}}(\nu_l) \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}(\nu_k) \hat{\boldsymbol{\mu}}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_l) \} \frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_k) \}}$ ,  $l, k = 1, \dots, \rho$ . Note that  $\boldsymbol{\Pi}$  is positive semi-definite. This can be shown by rewriting  $\boldsymbol{\Pi}$  in the form  $\boldsymbol{\Omega}^T \boldsymbol{\Omega}$  where

$$\boldsymbol{\Omega} = \begin{bmatrix} \frac{\bar{\boldsymbol{\Sigma}}^{1/2} \bar{\mathbf{Q}}(\nu_1) \hat{\boldsymbol{\mu}}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_1) \}} \cdots \frac{\bar{\boldsymbol{\Sigma}}^{1/2} \bar{\mathbf{Q}}(\nu_p) \hat{\boldsymbol{\mu}}}{\frac{1}{n-2} \text{tr} \{ \bar{\boldsymbol{\Sigma}} \bar{\mathbf{Q}}^2(\nu_p) \}} \end{bmatrix}.$$

Then  $\mathbf{a}^T \boldsymbol{\Omega}^T \boldsymbol{\Omega} \mathbf{a} = \|\boldsymbol{\Omega} \mathbf{a}\|_2^2 \geq 0$ ,  $\forall \mathbf{a} \neq \mathbf{0}$ . This, however, does not guarantee that  $\boldsymbol{\Pi}$  is positive-definite.

Using (B.18) and (B.19), the G-estimator of the probability of misclassification

(B.17) can be rewritten as

$$\sum_{i=0}^1 \hat{\pi}_i \Phi \left( (-1)^i \frac{\boldsymbol{\alpha}^T \left[ \frac{(-1)^{i+1}}{2} \mathbf{H}_1 \mathbf{b} + (-1)^i \mathbf{H}_\nu \mathbf{c}_i \right] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \right),$$

as presented in Theorem 4.4.1.

### B.1.2 Proof of Theorem 2

This section derives the  $\boldsymbol{\alpha}$  and  $\theta$  which jointly minimize the G-estimator of the probability of misclassification (4.15). The optimization problem is

$$\min_{\boldsymbol{\alpha}, \theta} f(\boldsymbol{\alpha}, \theta)$$

$\Downarrow$

$$\min_{\boldsymbol{\alpha}} \min_{\theta} f(\boldsymbol{\alpha}, \theta), \tag{B.20}$$

where

$$f(\boldsymbol{\alpha}, \theta) := \hat{\pi}_0 \Phi \left( \frac{\boldsymbol{\alpha}^T \left[ -\frac{1}{2} \mathbf{H}_1 \mathbf{b} + \mathbf{H}_\nu \mathbf{c}_0 \right] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \right) + \hat{\pi}_1 \Phi \left( -\frac{\boldsymbol{\alpha}^T \left[ \frac{1}{2} \mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}_1 \right] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \right)$$

We first find the stationary point  $\theta_0$  of the inner minimization of (B.20). We then show that this satisfies  $f''(\boldsymbol{\alpha}, \theta_0) > 0$  under a reasonable assumption on  $\boldsymbol{\alpha}$  that always holds in practice (otherwise the classifier performance is very bad and the class labels must be flipped, making the assumption hold anyway). Thus  $\theta_0$  is a local optimum and, since it is the only optimum, it is the global optimum,  $\theta^*$ , for a given  $\boldsymbol{\alpha}$ .

Writing out the CDF explicitly, we have

$$\min_{\boldsymbol{\alpha}} \min_{\theta} \left[ \frac{\hat{\pi}_0}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\boldsymbol{\alpha}^T [-\frac{1}{2}\mathbf{H}_1\mathbf{b} + \mathbf{H}_\nu\mathbf{c}_0] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \exp\left(-\frac{u^2}{2}\right) du + \frac{\hat{\pi}_1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\boldsymbol{\alpha}^T [\frac{1}{2}\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}_1] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \exp\left(-\frac{u^2}{2}\right) du \right].$$

Applying Leibniz's rule for differentiation under the integral sign, we have

$$\begin{aligned} f'(\boldsymbol{\alpha}, \theta) &= \frac{d \left[ \frac{\hat{\pi}_0}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\boldsymbol{\alpha}^T [-\frac{1}{2}\mathbf{H}_1\mathbf{b} + \mathbf{H}_\nu\mathbf{c}_0] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \exp\left(-\frac{u^2}{2}\right) du + \frac{\hat{\pi}_1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\boldsymbol{\alpha}^T [\frac{1}{2}\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}_1] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \exp\left(-\frac{u^2}{2}\right) du \right]}{d\theta} \\ &= \frac{\hat{\pi}_0}{\sqrt{2\pi}} \frac{1}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \exp\left(-\frac{1}{2} \frac{\left(\boldsymbol{\alpha}^T [-\frac{1}{2}\mathbf{H}_1\mathbf{b} + \mathbf{H}_\nu\mathbf{c}_0] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta\right)^2}{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}\right) - \\ &\quad \frac{\hat{\pi}_1}{\sqrt{2\pi}} \frac{1}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \exp\left(-\frac{1}{2} \frac{\left(\boldsymbol{\alpha}^T [\frac{1}{2}\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}_1] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta\right)^2}{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}\right). \end{aligned} \tag{B.21}$$

The stationary point,  $\theta_0$ , of (B.21) satisfies the equation

$$\hat{\pi}_0 \exp\left(-\frac{1}{2} \frac{\left(\boldsymbol{\alpha}^T [-\frac{1}{2}\mathbf{H}_1\mathbf{b} + \mathbf{H}_\nu\mathbf{c}_0] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta_0\right)^2}{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}\right) = \hat{\pi}_1 \exp\left(-\frac{1}{2} \frac{\left(\boldsymbol{\alpha}^T [\frac{1}{2}\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}_1] + \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta_0\right)^2}{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}\right). \tag{B.22}$$

Solving (B.22) for  $\theta_0$ , we have

$$\theta_0 = -\ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \frac{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} + \frac{\boldsymbol{\alpha}^T \mathbf{H}_\nu (\mathbf{c}_1 - \mathbf{c}_0)}{2}$$

The second derivative of  $f(\boldsymbol{\alpha}, \theta)$  is given by

$$\begin{aligned}
f''(\boldsymbol{\alpha}, \theta) = & -\frac{\hat{\pi}_0}{\sqrt{2\pi}} \frac{\boldsymbol{\alpha}^T [-\frac{1}{2}\mathbf{H}_1\mathbf{b} + \mathbf{H}_\nu\mathbf{c}_0] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{3/2}} \exp\left(-\frac{1}{2} \frac{(\boldsymbol{\alpha}^T [-\frac{1}{2}\mathbf{H}_1\mathbf{b} + \mathbf{H}_\nu\mathbf{c}_0] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta)^2}{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}\right) \\
& + \frac{\hat{\pi}_1}{\sqrt{2\pi}} \frac{\boldsymbol{\alpha}^T [\frac{1}{2}\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}_1] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta}{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{3/2}} \exp\left(-\frac{1}{2} \frac{(\boldsymbol{\alpha}^T [\frac{1}{2}\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}_1] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta)^2}{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}\right).
\end{aligned} \tag{B.23}$$

We have

$$\boldsymbol{\alpha}^T \left[-\frac{1}{2}\mathbf{H}_1\mathbf{b} + \mathbf{H}_\nu\mathbf{c}_0\right] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta_0 = -\frac{1}{2}\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}] + \frac{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}$$

and

$$\boldsymbol{\alpha}^T \left[\frac{1}{2}\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}_1\right] + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta_0 = \frac{1}{2}\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}] + \frac{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0},$$

where  $c := c_0 + c_1$ . Thus, substituting  $\theta_0$  into (B.23), we obtain

$$\begin{aligned}
f''(\boldsymbol{\alpha}, \theta_0) = & \frac{\hat{\pi}_0}{\sqrt{2\pi}} \frac{1}{2} \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]}{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{3/2}} \exp\left(-\frac{1}{2} \left(-\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) \\
& - \frac{\hat{\pi}_0}{\sqrt{2\pi}} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{-1/2}}{(\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}])} \exp\left(-\frac{1}{2} \left(-\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) \\
& + \frac{\hat{\pi}_1}{\sqrt{2\pi}} \frac{1}{2} \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]}{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{3/2}} \exp\left(-\frac{1}{2} \left(\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) \\
& + \frac{\hat{\pi}_1}{\sqrt{2\pi}} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{-1/2}}{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]} \exp\left(-\frac{1}{2} \left(\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) \\
= & \frac{1}{\sqrt{2\pi}} \frac{1}{2} \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]}{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{3/2}} \left[ \hat{\pi}_0 \exp\left(-\frac{1}{2} \left(-\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) + \hat{\pi}_1 \exp\left(-\frac{1}{2} \left(\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) \right] \\
& + \frac{1}{\sqrt{2\pi}} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{(\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha})^{-1/2}}{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]} \left[ -\hat{\pi}_0 \exp\left(-\frac{1}{2} \left(-\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) + \hat{\pi}_1 \exp\left(-\frac{1}{2} \left(\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \frac{1}{\tau}\right)^2\right) \right] \tag{B.24}
\end{aligned}$$

where  $\tau := \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}]}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \mathbf{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}}$ . The first term of the last line of (B.24) is clearly positive assuming that  $\boldsymbol{\alpha}^T [\mathbf{H}_1\mathbf{b} - \mathbf{H}_\nu\mathbf{c}] > 0$ , which holds for any reasonable  $\boldsymbol{\alpha}$  as having this quantity be negative would result in an error greater than 0.5 in which case the sign would have to be flipped anyway. Note that this quantity corresponds to the difference between the numerator estimators for the different

classes. Following from that,  $\boldsymbol{\alpha}^T [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}] = 0$  would not make sense either as that would mean the discriminant mean is the same for both classes. Assumption (c) ensures that this is not the case. Without loss of generality, we may assume that the majority class is  $\mathcal{C}_0$  so that  $\hat{\pi}_0 > \hat{\pi}_1$ . Then the last term of (B.24) is non-negative since

$$\exp\left(-\frac{1}{2}\left(-\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\frac{1}{\tau}\right)^2\right) \geq \exp\left(-\frac{1}{2}\left(\frac{1}{2}\tau + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\frac{1}{\tau}\right)^2\right)$$

and  $\ln\frac{\hat{\pi}_1}{\hat{\pi}_0} \leq 0$ .

Since  $f''(\boldsymbol{\alpha}, \theta_0) > 0$ ,  $\theta_0$  is a local minimum. Since it is the only minimum of  $f(\boldsymbol{\alpha}, \theta)$  over  $\mathbb{R}$ , it is a global minimum which we denote by  $\theta^*$ . Thus,

$$\theta^* = -\ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \frac{\boldsymbol{\alpha}^T \mathbf{H}_\nu \Pi \mathbf{H}_\nu^T \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \frac{\boldsymbol{\alpha}^T \mathbf{H}_\nu (\mathbf{c}_1 - \mathbf{c}_0)}{2}. \quad (\text{B.25})$$

Now we can proceed to the minimization over  $\boldsymbol{\alpha}$ . Substituting  $\theta^*$  back into (B.20), we have

$$\min_{\boldsymbol{\alpha}} \left[ \hat{\pi}_0 \Phi\left(-\frac{1}{2}\tau + \frac{1}{\tau} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \hat{\pi}_1 \Phi\left(-\frac{1}{2}\tau - \frac{1}{\tau} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) \right], \quad (\text{B.26})$$

Before solving the optimization problem (B.26), we note that we now have the G-estimator of the probability of misclassification of an LDA base classifier for any rotationally-invariant scheme (that satisfies the analyticity requirement (c)) with its corresponding optimal intercept, i.e.,

$$\hat{\pi}_0 \Phi\left(-\frac{1}{2}\tau + \frac{1}{\tau} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \hat{\pi}_1 \Phi\left(-\frac{1}{2}\tau - \frac{1}{\tau} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) \asymp \sum_{i=0}^1 \pi_i \Phi\left((-1)^i \frac{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \left(\boldsymbol{\mu}_i - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2}\right) + \ln\frac{\hat{\pi}_1}{\hat{\pi}_0} + \theta^*}{\sqrt{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \boldsymbol{\Sigma} \tilde{\boldsymbol{\Sigma}}_{\text{WS}} \hat{\boldsymbol{\mu}}}}\right).$$

We can solve (B.26) following steps similar to those in Appendix C of (Sifaou et al., 2020). Firstly, denote

$$g(\tau) := \hat{\pi}_0 \Phi\left(-\frac{1}{2}\tau + \frac{1}{\tau} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right) + \hat{\pi}_1 \Phi\left(-\frac{1}{2}\tau - \frac{1}{\tau} \ln\frac{\hat{\pi}_1}{\hat{\pi}_0}\right).$$

Then by differentiating under the integral, we obtain

$$\begin{aligned} \frac{dg(\tau)}{d\tau} &= \frac{\hat{\pi}_0}{\sqrt{2\pi}} \left( -\frac{1}{2} - \frac{1}{\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\frac{\left( -\frac{1}{2}\tau + \frac{1}{\tau} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2}{2} \right) \\ &\quad + \frac{\hat{\pi}_1}{\sqrt{2\pi}} \left( -\frac{1}{2} + \frac{1}{\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\frac{\left( -\frac{1}{2}\tau - \frac{1}{\tau} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2}{2} \right). \end{aligned}$$

Now multiplying both sides by  $\frac{1}{\hat{\pi}_0} \exp \left( -\frac{\left( -\frac{1}{2}\tau + \frac{1}{\tau} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2}{2} \right)$ , we have

$$\begin{aligned} \frac{1}{\hat{\pi}_0} \exp \left( -\frac{\left( -\frac{1}{2}\tau + \frac{1}{\tau} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2}{2} \right) \frac{dg(\tau)}{d\tau} &= \frac{1}{\sqrt{2\pi}} \left( -\frac{1}{2} - \frac{1}{\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\left( -\frac{1}{2}\tau + \frac{1}{\tau} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 \right) \\ &\quad + \frac{1}{\sqrt{2\pi}} \frac{\hat{\pi}_1}{\hat{\pi}_0} \left( -\frac{1}{2} + \frac{1}{\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\frac{1}{2} \left[ \left( -\frac{1}{2}\tau - \frac{1}{\tau} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 + \left( -\frac{1}{2}\tau + \frac{1}{\tau} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 \right] \right) \\ &= \frac{1}{\sqrt{2\pi}} \left( -\frac{1}{2} - \frac{1}{\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\frac{\tau^2}{4} - \frac{1}{\tau^2} \left( \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 \right) \exp \left( \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \\ &\quad + \frac{1}{\sqrt{2\pi}} \frac{\hat{\pi}_1}{\hat{\pi}_0} \left( -\frac{1}{2} + \frac{1}{\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\frac{\tau^2}{4} - \frac{1}{\tau^2} \left( \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 \right) \\ &= \frac{1}{\sqrt{2\pi}} \frac{\hat{\pi}_1}{\hat{\pi}_0} \left( -1 - \frac{1}{2\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\frac{\tau^2}{4} - \frac{1}{\tau^2} \left( \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 \right) \\ &\quad + \frac{1}{\sqrt{2\pi}} \frac{\hat{\pi}_1}{\hat{\pi}_0} \left( -1 + \frac{1}{2\tau^2} \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right) \exp \left( -\frac{\tau^2}{4} - \frac{1}{\tau^2} \left( \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 \right) \\ &= -\sqrt{\frac{2}{\pi}} \frac{\hat{\pi}_1}{\hat{\pi}_0} \exp \left( -\frac{\tau^2}{4} - \frac{1}{\tau^2} \left( \ln \frac{\hat{\pi}_1}{\hat{\pi}_0} \right)^2 \right) \\ &< 0. \end{aligned}$$

So  $g(\tau)$  is a decreasing function of  $\tau$ . Therefore the minimum of  $g(\tau)$  is obtained by maximizing  $\tau$  which is a function of  $\alpha$ . We therefore maximize  $\tau$  over  $\alpha$  to

obtain  $\boldsymbol{\alpha}^*$  as the solution to the following optimization problem

$$\boldsymbol{\alpha}^* = \operatorname{argmax}_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}}.$$

We can eigendecompose  $\mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T$  as

$$\begin{aligned} \mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T &= \mathbf{A} \mathbf{C} \mathbf{A}^T \\ &= \mathbf{A} \mathbf{C}^{1/2} \mathbf{C}^{1/2} \mathbf{A}^T \\ &= \mathbf{B} \mathbf{B}^T. \end{aligned}$$

Note that  $\mathbf{B}$  is invertible when  $\mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T$  is invertible. This is stated as an assumption in Theorem 4.4.1.

Now let  $u := \|\mathbf{B}^T \boldsymbol{\alpha}\|_2$  and  $\bar{\boldsymbol{\alpha}} := \frac{\mathbf{B}^T \boldsymbol{\alpha}}{u}$ . Then

$$\begin{aligned} \boldsymbol{\alpha}^* &= \operatorname{argmax}_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]}{\sqrt{\boldsymbol{\alpha}^T \mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T \boldsymbol{\alpha}}} \\ &= \operatorname{argmax}_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]}{\sqrt{\boldsymbol{\alpha}^T \mathbf{B} \mathbf{B}^T \boldsymbol{\alpha}}} \\ &= \operatorname{argmax}_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^T [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]}{u} \\ &= \operatorname{argmax}_{\|\boldsymbol{\alpha}\|_2=1, u>0} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]^T \mathbf{B}^{-T} \bar{\boldsymbol{\alpha}} \\ &= \operatorname{argmax}_{\|\bar{\boldsymbol{\alpha}}\|_2=1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]^T \mathbf{B}^{-T} \bar{\boldsymbol{\alpha}}, \quad u > 0, \end{aligned}$$

where the last line follows from the fact that the objective function is not a function of  $u$ . So,

$$\bar{\boldsymbol{\alpha}}^* = \frac{\mathbf{B}^{-1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]}{\|\mathbf{B}^{-1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]\|_2}$$

and

$$\boldsymbol{\alpha}^* = \frac{(\mathbf{B} \mathbf{B}^T)^{-1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]}{\|\mathbf{B}^{-1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]\|_2} u, \quad u > 0. \quad (\text{B.27})$$

The denominator of (B.27) can be included in the factor  $u$  so that we can rewrite

$\boldsymbol{\alpha}^*$  as follows without any reference to  $\mathbf{B}$ :

$$\boldsymbol{\alpha}^* = (\mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T)^{-1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}] u, \quad u > 0.$$

Now, for verification, let us go back and check the condition  $\boldsymbol{\alpha}^T [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}] \geq 0$  we assumed in order to ensure the convexity of the objective function when deriving the optimal intercept.

$$\begin{aligned} \boldsymbol{\alpha}^{*T} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}] &= [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}]^T (\mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T)^{-1} [\mathbf{H}_1 \mathbf{b} - \mathbf{H}_\nu \mathbf{c}] u, \quad u > 0, \\ &> 0, \end{aligned}$$

where the last line follows from the fact that  $(\mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T)^{-1}$  is positive definite by the assumption that  $\liminf_p \lambda_{\min}(\mathbf{H}_\nu \boldsymbol{\Pi} \mathbf{H}_\nu^T) > 0$  stated in Theorem 4.4.1.

## B.2 Other proofs

### B.2.1 Asymptotic form of the Marzetta precision matrix estimator shrinkage

We are interested in knowing the explicit function which acts on the sample eigenvalues for the Marzetta estimator of the precision matrix, `invcov`. In our past work, (Niyazi et al., 2020b), we found that, asymptotically and within the context of the discriminant-averaging RP-LDA ensemble classifier of (Durrant and Kabán, 2015), the shrinkage is a regularization of the form

$$d_j^{\text{disc-avg}} = \begin{cases} \gamma(d), & \text{if } l_j = 0 \\ \frac{1}{l_j + \frac{1}{\gamma(d)}}, & \text{otherwise,} \end{cases} \quad (\text{B.28})$$

for  $j = 1, \dots, p$ , where  $\gamma(d)$  is a regularization parameter, which, for fixed  $n$ ,  $p$ , and  $\hat{\boldsymbol{\Sigma}}$ , varies with the projection dimension  $d$ . This regularization acts uniformly on each sample eigenvalue  $d_i$  of  $\hat{\boldsymbol{\Sigma}}$ . More specifically,  $\gamma(d)$  is the root of the

monotonically decreasing function

$$f(x) = 1 - \frac{1}{d} \text{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{x} \mathbf{I}_p \right)^{-1} \right\} \quad (\text{B.29})$$

over  $x > 0$ . Also note that this work assumed Gaussian random projections with i.i.d. zero-mean entries each having variance  $1/d$ .

Outside of the context of the LDA probability of misclassification, an exact characterization of the modified eigenvalues of  $\text{invcov}$  is provided in (Marzetta et al., 2011) for random projections which are random unitary matrices drawn from Haar measure. Additionally, the data is assumed to be zero-mean Gaussian, resulting in  $n$  non-zero sample eigenvalues when  $\hat{\Sigma}$  is rank deficient. This result shows that  $\text{invcov}$  shifts all zero eigenvalues of the sample covariance matrix to a constant,  $\kappa$ , while it modifies the non-zero eigenvalues non-trivially to eigenvalues  $\chi_1, \dots, \chi_n$ . However, the exact form of the shrinkage is not explicit in these results because the expressions for  $\kappa$  and  $\{\chi_k\}_{k=1}^n$ , are intractable.

We resort to working with asymptotic relations for  $\kappa$  and  $\{\chi_k\}_{k=1}^n$ , also provided in (Marzetta et al., 2011), in order to derive an explicit expression for the shrinkage outside of the context of the LDA probability of misclassification. The resulting form of shrinkage is directly comparable to (B.28) derived under different assumptions because (Marzetta et al., 2011) only uses the unitary assumption to prove the closed form of the expressions for  $\kappa$  and  $\{\chi_k\}_{k=1}^n$ , while the proofs of the structure of the shrinkage (being constant for zero sample eigenvalues and non-trivial for non-zero sample eigenvalues) and the asymptotic relations which are used to derive the shrinkage in what follows make use of the fact that the unitary ensemble can be equivalently expressed as a Gaussian ensemble. We thus proceed with the proof as if we had assumed a Gaussian ensemble to begin with.

The asymptotic relation from (Marzetta et al., 2011) involving  $\kappa$  is

$$\frac{n-d}{n} \asymp \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + \kappa l_i} \quad (\text{B.30})$$

and the asymptotic relation from (Marzetta et al., 2011) involving  $\chi_k$  is

$$\chi_k \asymp \frac{\partial \kappa}{\partial l_k} \sum_{i=1}^n \frac{l_i}{1 + \kappa l_i} + \frac{\kappa}{1 + \kappa l_k} - \frac{d}{\kappa} \frac{\partial \kappa}{\partial l_k}. \quad (\text{B.31})$$

Taking the partial derivative of (B.30) with respect to  $l_k$ , we have

$$\begin{aligned} 0 &\asymp \frac{1}{n} \sum_{i=1}^n \frac{\partial[(1 + \kappa l_i)^{-1}]}{\partial l_k} \\ &= \frac{1}{n} \sum_{i \neq k} \frac{\partial[(1 + \kappa d_i)^{-1}]}{\partial l_k} + \frac{1}{n} \frac{\partial[(1 + \kappa l_k)^{-1}]}{\partial l_k} \\ &= -\frac{1}{n} \sum_{i \neq k} \frac{\frac{\partial \kappa}{\partial l_k} l_i}{(1 + \kappa l_i)^2} - \frac{1}{n} \frac{\frac{\partial \kappa}{\partial l_k} l_k + \kappa}{(1 + \kappa l_k)^2} \\ &= -\frac{1}{n} \sum_{i=1}^n \frac{\frac{\partial \kappa}{\partial l_k} l_i}{(1 + \kappa l_i)^2} - \frac{1}{n} \frac{\kappa}{(1 + \kappa l_k)^2}, \end{aligned}$$

which yields

$$\frac{\partial \kappa}{\partial l_k} \asymp -\frac{1}{n} \frac{\frac{\kappa}{(1 + \kappa l_k)^2}}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1 + \kappa l_i)^2}}. \quad (\text{B.32})$$

Note that  $\frac{\partial \kappa}{\partial l_k}$  is of order  $1/n$ . Substituting (B.32) into (B.31), we obtain

$$\begin{aligned} \chi_k &\asymp -\frac{\kappa}{(1 + \kappa l_k)^2} \frac{\frac{1}{n} \sum_{i=1}^n \frac{d_i}{1 + \kappa l_i}}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1 + \kappa l_i)^2}} + \frac{\kappa}{1 + \kappa l_k} + \frac{d}{n} \frac{1}{(1 + \kappa l_k)^2} \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1 + \kappa l_i)^2}} \\ &= \frac{1/\kappa^2}{(l_k + 1/\kappa)^2} \left[ -\kappa \frac{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{1 + \kappa l_i}}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1 + \kappa l_i)^2}} + \frac{d}{n} \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1 + \kappa l_i)^2}} \right] + \frac{1}{l_k + 1/\kappa} \\ &= \frac{C}{(l_k + 1/\kappa)^2} + \frac{1}{l_k + 1/\kappa}, \end{aligned}$$

where  $C := \frac{1}{\kappa^2} \left[ -\kappa \frac{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{1 + \kappa l_i}}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1 + \kappa l_i)^2}} + \frac{d}{n} \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1 + \kappa l_i)^2}} \right]$ . We can show that  $C \asymp 0$  as

follows

$$\begin{aligned}
C &= \frac{1}{\kappa^2} \left[ -\kappa \frac{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{1+\kappa l_i}}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1+\kappa l_i)^2}} + \frac{d}{n} \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1+\kappa l_i)^2}} \right] \\
&= \frac{1}{\mu^2} \left[ \frac{d/n - \left( \frac{1}{n} \sum_{i=1}^n \frac{\kappa l_i}{1+\kappa l_i} - 1 + 1 \right)}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1+\kappa l_i)^2}} \right] \\
&= \frac{1}{\kappa^2} \left[ \frac{d/n + \frac{1}{n} \sum_{i=1}^n \frac{1}{1+\kappa l_i} - 1}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1+\kappa l_i)^2}} \right] \\
&\asymp \frac{1}{\kappa^2} \left[ \frac{d/n + \frac{n-d}{n} - 1}{\frac{1}{n} \sum_{i=1}^n \frac{l_i}{(1+\kappa l_i)^2}} \right] \\
&= 0,
\end{aligned}$$

where the second-to-last line uses the asymptotic relation in (B.30).

We are left with only the term  $\chi_k \asymp \frac{1}{l_k+1/\kappa}$  so that overall the form of the shrinkage of invcov is asymptotically a regularization uniformly across all sample eigenvalues by  $1/\kappa$ , i.e.,

$$d_j^{\text{invcov}} = \begin{cases} \kappa & \text{if } l_j = 0 \\ \frac{1}{d_j+1/\kappa}, & \text{otherwise.} \end{cases} \quad (\text{B.33})$$

This suggests that  $\gamma(d)$  in (B.28) plays the role of  $\kappa$  in (B.33). In fact, we can show that the  $\gamma(d) = \kappa$  by showing that the roots of (B.29) and (B.30) are equivalent. First, (B.30) suggests that, asymptotically,  $\kappa$  is the root of the function

$$g(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1+x d_i} + \frac{d}{n} - 1$$

over  $x > 0$ . We can show that the root of  $f(x)$  is equal to the root of  $g(x)$  as

follows:

$$\begin{aligned}
f(x) &= 1 - \frac{1}{d} \operatorname{tr} \left\{ \hat{\Sigma} \left( \hat{\Sigma} + \frac{1}{x} \mathbf{I}_p \right)^{-1} \right\} \\
&= 1 - \frac{1}{d} \operatorname{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{x} \mathbf{I}_p - \frac{1}{x} \mathbf{I}_p \right) \left( \hat{\Sigma} + \frac{1}{x} \mathbf{I}_p \right)^{-1} \right\} \\
&= 1 - \frac{p}{d} + \frac{1}{x} \frac{1}{d} \operatorname{tr} \left\{ \left( \hat{\Sigma} + \frac{1}{x} \mathbf{I}_p \right)^{-1} \right\} \\
&= 1 - \frac{p}{d} + \frac{1}{d} \sum_{i=1}^p \frac{1/x}{d_i + 1/x} \\
&= 1 - \frac{p}{d} + \frac{1}{d} \sum_{i=1}^n \frac{1}{1 + x d_i} + \frac{p-n}{d} \\
&= \frac{1}{d} \sum_{i=1}^n \frac{1}{1 + x d_i} + 1 - \frac{n}{d}
\end{aligned} \tag{B.34}$$

where the second-to-last line uses the fact that there are  $n$  non-zero sample eigenvalues under Marzetta's assumption of zero-mean data. Since scaling  $f(x)$  by a constant doesn't alter its root, we can multiply (B.34) by  $d/n$  to show that the roots of  $f(x)$  and  $g(x)$  are equivalent, and therefore  $\gamma(d) = \kappa$ .

## B.2.2 Bayes shrinkage is rotationally-invariant

As shown in Section 4.3.2, the Bayes shrinkage is

$$d_j^{\text{Bayes}} := \frac{\mathbf{u}_j^T \Sigma^{-1} \boldsymbol{\mu}}{\mathbf{u}_j^T \boldsymbol{\mu}}, \quad j = 1, \dots, p,$$

and, in fact, any shrinkage proportional to this achieves the Bayes error for two-class Gaussian data with a common covariance. However, this shrinkage is not a function of the sample eigenvalues, but of the sample eigenvectors, true means, and true covariances. Does such a shrinkage retain the property of rotational-invariance? In this case it does, as we show in the following.

We write the Bayes shrinkage using the notation

$$d_j^{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1) := \phi(\mathbf{u}_j, \Sigma, \boldsymbol{\mu}) = \frac{\mathbf{u}_j^T \Sigma^{-1} \boldsymbol{\mu}}{\mathbf{u}_j^T \boldsymbol{\mu}}, \quad j = 1, \dots, p,$$

with the arguments  $\mathbf{X}_0$  and  $\mathbf{X}_1$  included to distinguish between the case where the shrinkage depends on unrotated data and the case where it depends on rotated data. Let

$$\tilde{\Sigma}_{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1) := \mathbf{U} \mathbf{D}_{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1) \mathbf{U}^T,$$

where

$$\mathbf{D}_{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1) = \text{diag}\left(d_1^{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1), \dots, d_p^{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1)\right),$$

denote the rotationally-invariant estimator corresponding to the Bayes shrinkage. To check rotational-invariance, we rotate the data by an orthogonal matrix  $\mathbf{W}$ . The eigenvectors of the sample covariance based on the rotated data are easily verified to be  $\mathbf{W}\mathbf{U}$ . The true statistics corresponding to the rotated data are easily shown to be  $\mathbf{W}\Sigma\mathbf{W}^T$  and  $\mathbf{W}\boldsymbol{\mu}$ . Then it is easy to see that  $\phi(\mathbf{W}\mathbf{u}_j, \mathbf{W}\Sigma\mathbf{W}^T, \mathbf{W}\boldsymbol{\mu}) = \phi(\mathbf{u}_j, \Sigma, \boldsymbol{\mu})$  and therefore

$$\begin{aligned} \tilde{\Sigma}_{\text{Bayes}}(\mathbf{W}\mathbf{X}_0, \mathbf{W}\mathbf{X}_1) &= \mathbf{W}\mathbf{U}\mathbf{D}_{\text{Bayes}}(\mathbf{W}\mathbf{X}_0, \mathbf{W}\mathbf{X}_1) \mathbf{U}^T \mathbf{W}^T \\ &= \mathbf{W}\mathbf{U}\mathbf{D}_{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1) \mathbf{U}^T \mathbf{W}^T \\ &= \mathbf{W}\tilde{\Sigma}_{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1) \mathbf{W}^T. \end{aligned}$$

Thus,  $\tilde{\Sigma}_{\text{Bayes}}(\mathbf{X}_0, \mathbf{X}_1)$  is rotationally-invariant.

## Appendix C

### Proofs for Chapter 5

#### C.1 Analysis of the projected test point in the case of unknown means

##### C.1.1 Common covariances

As in the derivation of Section 5.3.1, assume  $\mathbf{x} \in \mathcal{C}_i$ , where  $i$  is either 0 or 1 and assume the two classes have a common covariance matrix  $\Sigma$ . Then  $\mathbf{x}|\mathbf{x} \in \mathcal{C}_i \sim \boldsymbol{\mu}_i + \Sigma^{1/2}\mathbf{z}$  where  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ .

Using the fact that  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \frac{\Sigma^{1/2}\mathbf{Z}_i\mathbf{1}}{n_i}$  for some  $\mathbf{Z}_i$  with  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  columns,  $i = 0, 1$ ,  $\hat{\mathbf{x}}$  can be expressed as

$$\hat{\mathbf{x}}|\mathbf{x} \in \mathcal{C}_i = \frac{(-1)^{i+1}}{2}\boldsymbol{\mu} + \Sigma^{1/2}\mathbf{z} - \frac{1}{2}\frac{\Sigma^{1/2}\mathbf{Z}_0\mathbf{1}}{n_0} - \frac{1}{2}\frac{\Sigma^{1/2}\mathbf{Z}_1\mathbf{1}}{n_1}. \quad (\text{C.1})$$

The first term in (5.11) can then be rewritten as

$$\frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \hat{\mathbf{x}}|\mathbf{x} \in \mathcal{C}_i = \underbrace{(-1)^{i+1} \frac{1}{2} \frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \boldsymbol{\mu}}_{I_1(\text{information})} + \overbrace{\frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \left( \Sigma^{1/2}\mathbf{z} - \frac{1}{2}\frac{\Sigma^{1/2}\mathbf{Z}_0\mathbf{1}}{n_0} - \frac{1}{2}\frac{\Sigma^{1/2}\mathbf{Z}_1\mathbf{1}}{n_1} \right)}^{N_1(\text{noise})}.$$

Note that the noise here is due to both the common covariance between the classes (this is the test point noise) as well as estimation noise from the sample

means. Similarly, the second term can be expressed using (C.1) as

$$\mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \hat{\mathbf{x}} | \mathbf{x} \in \mathcal{C}_i = \underbrace{\frac{(-1)^{i+1}}{2} \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \boldsymbol{\mu}}_{I_2(\text{information})} + \underbrace{\mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \left( \boldsymbol{\Sigma}^{1/2} \mathbf{z} - \frac{1}{2} \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} - \frac{1}{2} \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1} \right)}_{N_2(\text{noise})}.$$

Alternatively, expressing (C.1) in terms of  $\hat{\boldsymbol{\mu}}$  rather than  $\boldsymbol{\mu}$  so that the orthogonal projection can be put to use yields a similar result. By using the fact that  $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} - \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1}$ , (C.1) can be expressed as

$$\hat{\mathbf{x}} | \mathbf{x} \in \mathcal{C}_i = \frac{(-1)^{i+1}}{2} \hat{\boldsymbol{\mu}} + \boldsymbol{\Sigma}^{1/2} \mathbf{z} - \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$$

and the second term in (5.11) as

$$\begin{aligned} \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \hat{\mathbf{x}} | \mathbf{x} \in \mathcal{C}_i &= \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \left( \frac{(-1)^{i+1}}{2} \hat{\boldsymbol{\mu}} + \boldsymbol{\Sigma}^{1/2} \mathbf{z} - \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i} \right) \\ &= \underbrace{\mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \boldsymbol{\Sigma}^{1/2} \mathbf{z}}_{N_2(\text{noise})} - \underbrace{\mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \left( \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i} \right)}_{I_2(\text{information})}. \end{aligned}$$

From this perspective, the information in the test point combines with the sample estimation noise in the term  $\frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$ . Note that even if we were to have equal samples so that  $n_0 = n_1$ , we would still be able to discriminate the class of the test point through  $\mathbf{Z}_i$ . This is not immediately obvious as  $\mathbf{Z}_i$ ,  $i = 0, 1$ , both have the same distribution, but can be observed asymptotically by computing DEs.

### C.1.2 Distinct covariances

An analogous result to that of the previous section can be derived in the case when the class covariance matrices are distinct. In this case,  $\mathbf{x} \sim \boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i^{1/2} \mathbf{z}$  where  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Using the fact that  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$  for some  $\mathbf{Z}_i$  with

$\mathcal{N}(\mathbf{0}, \mathbf{I})$  columns,  $i = 0, 1$ ,

$$\hat{\mathbf{x}}|\mathbf{x} \in \mathcal{C}_i = \frac{(-1)^{i+1}}{2} \boldsymbol{\mu} + \boldsymbol{\Sigma}_i^{1/2} \mathbf{z} - \frac{1}{2} \frac{\boldsymbol{\Sigma}_0^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} - \frac{1}{2} \frac{\boldsymbol{\Sigma}_1^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1}. \quad (\text{C.2})$$

The first term in (5.11) can then be rewritten as

$$\frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \hat{\mathbf{x}}|\mathbf{x} \in \mathcal{C}_i = \underbrace{(-1)^{i+1} \frac{1}{2} \frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \boldsymbol{\mu} + \frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_i^{1/2} \mathbf{z}}_{I_1(\text{information})} - \frac{1}{2} \underbrace{\frac{\mathbf{w}^T \hat{\boldsymbol{\mu}}}{\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}}} \hat{\boldsymbol{\mu}}^T \left( \frac{\boldsymbol{\Sigma}_0^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} + \frac{\boldsymbol{\Sigma}_1^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1} \right)}_{N_1(\text{noise 1})}$$

Note here that the noise is due only to estimation noise from the sample means, since the differing covariances between the two classes are informative.

Substituting (C.2) directly into the second term in (5.11) gives

$$\mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \hat{\mathbf{x}}|\mathbf{x} \in \mathcal{C}_i = \underbrace{\frac{(-1)^{i+1}}{2} \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \boldsymbol{\mu} + \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \boldsymbol{\Sigma}_i^{1/2} \mathbf{z}}_{I_2(\text{information})} - \frac{1}{2} \underbrace{\mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \left( \frac{\boldsymbol{\Sigma}_0^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} + \frac{\boldsymbol{\Sigma}_1^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1} \right)}_{N_2(\text{noise})}$$

Alternatively, expressing (C.2) in terms of  $\hat{\boldsymbol{\mu}}$  reveals the second term in (5.11) to be purely information. By using the fact that  $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} + \frac{\boldsymbol{\Sigma}_0^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} - \frac{\boldsymbol{\Sigma}_1^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1}$ ,

$$\hat{\mathbf{x}}|\mathbf{x} \in \mathcal{C}_i = \frac{(-1)^{i+1}}{2} \hat{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_i^{1/2} \mathbf{z} - \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i},$$

and the second term in (5.11) becomes

$$\begin{aligned} \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \hat{\mathbf{x}}|\mathbf{x} \in \mathcal{C}_i &= \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \left( \frac{(-1)^{i+1}}{2} \hat{\boldsymbol{\mu}} + \boldsymbol{\Sigma}_i^{1/2} \mathbf{z} - \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i} \right) \\ &= \underbrace{\mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \boldsymbol{\Sigma}_i^{1/2} \mathbf{z} - \mathbf{w}^T \mathbf{P}_{\hat{\boldsymbol{\mu}}} \left( \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i} \right)}_{I_2(\text{information})}. \end{aligned}$$

## C.2 Derivation of the deterministic equivalent of the probability of misclassification

Deriving  $\bar{\varepsilon}$  reduces to deriving the DEs  $\bar{m}_0$ ,  $\bar{m}_1$ ,  $\bar{\sigma}_0^2$ , and  $\bar{\sigma}_1^2$ , which reduces to deriving the DEs of the constituent quadratic forms of  $m_0$ ,  $m_1$ ,  $\sigma_0^2$ , and  $\sigma_1^2$ . This is the approach taken in what follows.

### C.2.1 Distinct covariances

The following proofs rely heavily on three main facts. First, under the distinct covariance assumption on the class distributions,  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$  for some  $\mathbf{Z}_i$  with  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  columns,  $i = 0, 1$ . This follows from expressing the data matrices  $\mathbf{X}_i$ ,  $i = 0, 1$  as  $\mathbf{X}_i = \boldsymbol{\mu}_i \mathbf{1}_{n_i}^T + \boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i$  for some  $\mathbf{Z}_i$ ,  $i = 0, 1$  with columns distributed as  $\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$ . Then  $\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \mathbf{X}_i \mathbf{1}_{n_i} = \boldsymbol{\mu}_i + \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$ .

Second, the sample means  $\boldsymbol{\mu}_0$  and  $\boldsymbol{\mu}_1$  are independent of the sample covariance  $\boldsymbol{\Sigma}$ . This can be shown by simply plugging in  $\mathbf{X}_i = \boldsymbol{\mu}_i \mathbf{1}_{n_i}^T + \boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i$  and  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$  into the corresponding formula for  $\hat{\boldsymbol{\Sigma}}_i$ . This yields

$$\hat{\boldsymbol{\Sigma}}_i = \frac{1}{n_i - 1} \boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \left( \mathbf{I}_{n_i} - \frac{\mathbf{1}_{n_i} \mathbf{1}_{n_i}^T}{n_i} \right) \mathbf{Z}_i^T \boldsymbol{\Sigma}_i^{1/2}. \quad (\text{C.3})$$

Since the terms  $\mathbf{Z}_i \mathbf{1}_{n_i}$  and  $\mathbf{Z}_i \left( \mathbf{I}_{n_i} - \frac{\mathbf{1}_{n_i} \mathbf{1}_{n_i}^T}{n_i} \right)$  are Gaussian and uncorrelated, due to the projection matrix  $\left( \mathbf{I}_{n_i} - \frac{\mathbf{1}_{n_i} \mathbf{1}_{n_i}^T}{n_i} \right)$ , they are independent. Thus,  $\hat{\boldsymbol{\mu}}_i$  and  $\hat{\boldsymbol{\Sigma}}_i$  are independent. Of course,  $\hat{\boldsymbol{\mu}}_i$  and  $\hat{\boldsymbol{\Sigma}}_i$  where  $i \neq j$  are also independent since  $\mathbf{Z}_0$  is independent of  $\mathbf{Z}_1$ . It follows that  $\hat{\boldsymbol{\Sigma}}$  which is a function of  $\hat{\boldsymbol{\Sigma}}_i$ ,  $i = 0, 1$ , is independent of  $\hat{\boldsymbol{\mu}}_i$ ,  $i = 0, 1$ .

Lastly,  $\hat{\boldsymbol{\Sigma}}$  can be expressed as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \boldsymbol{\Sigma}_0^{1/2} \bar{\mathbf{Z}}_0 \bar{\mathbf{Z}}_0^T \boldsymbol{\Sigma}_0^{1/2} + \frac{1}{n-2} \boldsymbol{\Sigma}_1^{1/2} \bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^T \boldsymbol{\Sigma}_1^{1/2}$$

for some  $\bar{\mathbf{Z}}_0 \in \mathbb{R}^{p \times (n_0-1)}$  and  $\bar{\mathbf{Z}}_1 \in \mathbb{R}^{p \times (n_1-1)}$ , both having columns distributed as

$\mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$ . Using (C.3),

$$\hat{\Sigma} = \frac{1}{n-2} \Sigma_0^{1/2} \mathbf{Z}_0 \left( \mathbf{I}_{n_0} - \frac{\mathbf{1}_{n_0} \mathbf{1}_{n_0}^T}{n_0} \right) \mathbf{Z}_0^T \Sigma_0^{1/2} + \frac{1}{n-2} \Sigma_1^{1/2} \mathbf{Z}_1 \left( \mathbf{I}_{n_1} - \frac{\mathbf{1}_{n_1} \mathbf{1}_{n_1}^T}{n_1} \right) \mathbf{Z}_1^T \Sigma_1^{1/2}.$$

Since the terms  $\frac{\mathbf{1}_{n_0} \mathbf{1}_{n_0}^T}{n_0}$  and  $\frac{\mathbf{1}_{n_1} \mathbf{1}_{n_1}^T}{n_1}$  each have one eigenvalue which is equal to 1 in both cases, their eigendecompositions can be represented as

$$\frac{\mathbf{1}_{n_0} \mathbf{1}_{n_0}^T}{n_0} = \mathbf{U}_0 \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \mathbf{U}_0^T \quad \text{and} \quad \frac{\mathbf{1}_{n_1} \mathbf{1}_{n_1}^T}{n_1} = \mathbf{U}_1 \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \mathbf{U}_1^T, \quad (\text{C.4})$$

where  $\mathbf{U}_0$  and  $\mathbf{U}_1$  have as their first columns the vectors  $\frac{\mathbf{1}_{n_0}}{\sqrt{n_0}}$  and  $\frac{\mathbf{1}_{n_1}}{\sqrt{n_1}}$  respectively.

By using these same bases to eigendecompose  $\mathbf{I}_{n_0}$  and  $\mathbf{I}_{n_1}$  in (C.4), we obtain

$$\begin{aligned} \hat{\Sigma} &= \frac{1}{n-2} \Sigma_0^{1/2} \mathbf{Z}_0 \mathbf{U}_0 \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \mathbf{U}_0^T \mathbf{Z}_0^T \Sigma_0^{1/2} + \frac{1}{n-2} \Sigma_1^{1/2} \mathbf{Z}_1 \mathbf{U}_1 \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \mathbf{U}_1^T \mathbf{Z}_1^T \Sigma_1^{1/2} \\ &\sim \frac{1}{n-2} \Sigma_0^{1/2} \mathbf{Z}_0 \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \mathbf{Z}_0^T \Sigma_0^{1/2} + \frac{1}{n-2} \Sigma_1^{1/2} \mathbf{Z}_1 \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}^T \mathbf{Z}_1^T \Sigma_1^{1/2} \\ &= \frac{1}{n-2} \Sigma_0^{1/2} \bar{\mathbf{Z}}_0 \bar{\mathbf{Z}}_0^T \Sigma_0^{1/2} + \frac{1}{n-2} \Sigma_1^{1/2} \bar{\mathbf{Z}}_1 \bar{\mathbf{Z}}_1^T \Sigma_1^{1/2}, \end{aligned}$$

where  $\bar{\mathbf{Z}}_0 \in \mathbb{R}^{p \times (n_0-1)}$  is the sub-matrix of  $\mathbf{Z}_0$  obtained by removing its first column and  $\bar{\mathbf{Z}}_1 \in \mathbb{R}^{p \times (n_1-1)}$  is the sub-matrix of  $\mathbf{Z}_1$  obtained by removing its first column.

Now we are ready to derive the DEs.

### Derivation of $\bar{m}_0$

The discriminant mean  $m_0$  can be expressed as

$$m_0 = (1 - \alpha)\rho\hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \alpha\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$$

Thus, the problem of deriving this DE can be further decomposed into deriving the following convergence statements

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \}$$

$$\hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \asymp -\frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \} - \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \} \right)$$

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \bar{\mathbf{Q}} \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \bar{\mathbf{Q}} \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \bar{\mathbf{Q}} \}$$

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \asymp -\frac{1}{2} \boldsymbol{\mu}^T \bar{\mathbf{Q}} \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \bar{\mathbf{Q}} \} - \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \bar{\mathbf{Q}} \} \right)$$

The first two convergence statements are derived by using the fact that  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \frac{\boldsymbol{\Sigma}_i^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$  for some  $\mathbf{Z}_i$  with  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  columns,  $i = 0, 1$  and taking the expectation.

The terms converge to their respective expectations according to Lemmas 17 and 19 in (Müller and Debbah, 2016). The third and fourth terms involve  $\hat{\boldsymbol{\Sigma}}$ . Since  $\hat{\boldsymbol{\mu}}$  and  $\hat{\boldsymbol{\Sigma}}$  are independent, the convergence can be split into stages.

For the third term, we first have the intermediate convergence result

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \hat{\boldsymbol{\Sigma}}^{-1} \},$$

and for the fourth term we have the intermediate convergence result

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \asymp -\frac{1}{2} \boldsymbol{\mu}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \} - \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \hat{\boldsymbol{\Sigma}}^{-1} \} \right)$$

each obtained by dealing with  $\hat{\boldsymbol{\mu}}$  as described above independently of  $\hat{\boldsymbol{\Sigma}}$ .

Next, we express  $\hat{\boldsymbol{\Sigma}} = \mathbf{W}\mathbf{W}^T$  where  $\mathbf{W} \in \mathbb{R}^{p \times (n-2)}$  is defined as

$$\mathbf{W} = \frac{1}{\sqrt{p}} \left[ \sqrt{\frac{p}{n-2}} \boldsymbol{\Sigma}_0^{1/2} \bar{\mathbf{Z}}_0 \quad \sqrt{\frac{p}{n-2}} \boldsymbol{\Sigma}_1^{1/2} \bar{\mathbf{Z}}_1 \right]$$

Now define  $\mathbf{Q}_\gamma = (\mathbf{W}\mathbf{W}^T - \gamma \mathbf{I}_p)^{-1}$ ,  $\gamma < 0$ . According to (Benaych-Georges and Couillet, 2016),

$$\mathbf{Q}_\gamma \leftrightarrow \bar{\mathbf{Q}}_\gamma,$$

where

$$\bar{\mathbf{Q}}_\gamma = -\frac{1}{\gamma} \left( \mathbf{I}_p + \frac{n_0 - 1}{n - 2} \delta(\gamma) \frac{p}{n - 2} \boldsymbol{\Sigma}_0 + \frac{n_1 - 1}{n - 2} \nu(\gamma) \frac{p}{n - 2} \boldsymbol{\Sigma}_1 \right)^{-1}, \quad (\text{C.5})$$

$$\frac{p}{n - 2} \delta(\gamma) = -\frac{1}{\gamma} \frac{1}{1 + \tilde{\delta}(\gamma)}, \quad (\text{C.6})$$

$$\tilde{\delta}(\gamma) = -\frac{1}{\gamma} \frac{1}{p} \text{tr} \left\{ \frac{p}{n - 2} \boldsymbol{\Sigma}_0 \left( \mathbf{I}_p + \frac{n_0 - 1}{n - 2} \delta(\gamma) \frac{p}{n - 2} \boldsymbol{\Sigma}_0 + \frac{n_1 - 1}{n - 2} \nu(\gamma) \frac{p}{n - 2} \boldsymbol{\Sigma}_1 \right)^{-1} \right\}, \quad (\text{C.7})$$

$$\frac{p}{n - 2} \nu(\gamma) = -\frac{1}{\gamma} \frac{1}{1 + \tilde{\nu}(\gamma)}, \quad (\text{C.8})$$

and

$$\tilde{\nu}(\gamma) = -\frac{1}{\gamma} \frac{1}{p} \text{tr} \left\{ \frac{p}{n - 2} \boldsymbol{\Sigma}_1 \left( \mathbf{I}_p + \frac{n_0 - 1}{n - 2} \delta(\gamma) \frac{p}{n - 2} \boldsymbol{\Sigma}_0 + \frac{n_1 - 1}{n - 2} \nu(\gamma) \frac{p}{n - 2} \boldsymbol{\Sigma}_1 \right)^{-1} \right\} \quad (\text{C.9})$$

The expressions we are working with can be expressed in this notation as

$$\lim_{\gamma \rightarrow 0} \hat{\boldsymbol{\mu}}^T \mathbf{Q}_\gamma \hat{\boldsymbol{\mu}}$$

and

$$\lim_{\gamma \rightarrow 0} \hat{\boldsymbol{\mu}}^T \mathbf{Q}_\gamma \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$$

and we want to derive the corresponding DEs by taking the limits

$$\lim_{n,p \rightarrow \infty} \lim_{\gamma \rightarrow 0} \hat{\boldsymbol{\mu}}^T \mathbf{Q}_\gamma \hat{\boldsymbol{\mu}}$$

$$\lim_{n,p \rightarrow \infty} \lim_{\gamma \rightarrow 0} \hat{\boldsymbol{\mu}}^T \mathbf{Q}_\gamma \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$$

The Moore-Osgood theorem allows the interchange of these limits. It is enough to show that the sequences  $\hat{\boldsymbol{\mu}}^T \mathbf{Q}_\gamma \hat{\boldsymbol{\mu}}$  and  $\hat{\boldsymbol{\mu}}^T \mathbf{Q}_\gamma \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$  converge uniformly. Since these sequences converge pointwise (this follows from convergence in probability), this can be shown by uniformly bounding their first derivative (Fischer, 2014).

We have

$$\begin{aligned} \frac{d [\hat{\boldsymbol{\mu}}^T \mathbf{Q}_\gamma \hat{\boldsymbol{\mu}}]}{d\gamma} &= \hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\Sigma}} - \gamma \mathbf{I}_p \right)^{-2} \hat{\boldsymbol{\mu}}^T \\ &\leq \|\hat{\boldsymbol{\mu}}\|_2^2 \left\| \left( \hat{\boldsymbol{\Sigma}} - \gamma \mathbf{I}_p \right)^{-1} \right\|_2^2 \\ &= \frac{\|\hat{\boldsymbol{\mu}}\|_2^2}{\lambda_{\min}\{\hat{\boldsymbol{\Sigma}}\} - \gamma} \\ &\leq \frac{\|\hat{\boldsymbol{\mu}}\|_2^2}{C}, \end{aligned}$$

where the last line follows from the result in (Kammoun and Alouini, 2014) which shows that  $\lambda_{\min}\{\hat{\boldsymbol{\Sigma}}\} > C$  for some constant  $C$  almost surely. Using the growth regime assumption (c) it can be shown that  $\|\hat{\boldsymbol{\mu}}\|_2^2$  is bounded. This completes the proof. The other term can be handled in a similar way.

We can now apply the result in (Benaych-Georges and Couillet, 2016) which manifests in equations (C.5), (C.6), (C.7), (C.8), and (C.9). We can then take the limit as  $\gamma \rightarrow 0$ .

Combining (C.6) and (C.7), we have

$$\frac{p}{n-2}\delta(\gamma) = -\frac{1}{\gamma} \frac{1}{1 - \frac{1}{\gamma} \frac{1}{p} \text{tr} \left\{ \frac{p}{n-2} \mathbf{\Sigma}_0 \left( \mathbf{I}_p + \frac{n_0-1}{n-2} \delta(\gamma) \frac{p}{n-2} \mathbf{\Sigma}_0 + \frac{n_1-1}{n-2} \nu(\gamma) \frac{p}{n-2} \mathbf{\Sigma}_1 \right)^{-1} \right\}}.$$

Combining (C.8) and (C.9), we have

$$\frac{p}{n-2}\nu(\gamma) = -\frac{1}{\gamma} \frac{1}{1 - \frac{1}{\gamma} \frac{1}{p} \text{tr} \left\{ \frac{p}{n-2} \mathbf{\Sigma}_1 \left( \mathbf{I}_p + \frac{n_0-1}{n-2} \delta(\gamma) \frac{p}{n-2} \mathbf{\Sigma}_0 + \frac{n_1-1}{n-2} \nu(\gamma) \frac{p}{n-2} \mathbf{\Sigma}_1 \right)^{-1} \right\}}.$$

These equations shows that  $\delta(\gamma)$  and  $\nu(\gamma)$  vary as  $\frac{1}{\gamma}$ , and so they diverge as  $\gamma \rightarrow 0$

Combining (C.7), (C.6), and (C.8), we have

$$\tilde{\delta}(\gamma) = \frac{1}{n-2} \text{tr} \left\{ \mathbf{\Sigma}_0 \left( -\gamma \mathbf{I}_p + \frac{n_0-1}{n-2} \frac{1}{1 + \tilde{\delta}(\gamma)} \mathbf{\Sigma}_0 + \frac{n_1-1}{n-2} \frac{1}{1 + \tilde{\nu}(\gamma)} \mathbf{\Sigma}_1 \right)^{-1} \right\}, \quad (\text{C.10})$$

Combining (C.9), (C.6), and (C.8), we have

$$\tilde{\nu}(\gamma) = \frac{1}{n-2} \text{tr} \left\{ \mathbf{\Sigma}_1 \left( -\gamma \mathbf{I}_p + \frac{n_0-1}{n-2} \frac{1}{1 + \tilde{\delta}(\gamma)} \mathbf{\Sigma}_0 + \frac{n_1-1}{n-2} \frac{1}{1 + \tilde{\nu}(\gamma)} \mathbf{\Sigma}_1 \right)^{-1} \right\}, \quad (\text{C.11})$$

This pair of equations does not pose problems as  $\gamma \rightarrow 0$ , therefore we work with  $\tilde{\delta}(\gamma)$  and  $\tilde{\nu}(\gamma)$ . Taking the limit as  $\gamma \rightarrow 0$ , (C.10) becomes

$$\tilde{\delta}(0) = \frac{1}{n-2} \text{tr} \left\{ \mathbf{\Sigma}_0 \left( \frac{n_0-1}{n-2} \frac{1}{1 + \tilde{\delta}(0)} \mathbf{\Sigma}_0 + \frac{n_1-1}{n-2} \frac{1}{1 + \tilde{\nu}(0)} \mathbf{\Sigma}_1 \right)^{-1} \right\},$$

and (C.11) becomes

$$\tilde{\nu}(0) = \frac{1}{n-2} \text{tr} \left\{ \mathbf{\Sigma}_1 \left( \frac{n_0-1}{n-2} \frac{1}{1 + \tilde{\delta}(0)} \mathbf{\Sigma}_0 + \frac{n_1-1}{n-2} \frac{1}{1 + \tilde{\nu}(0)} \mathbf{\Sigma}_1 \right)^{-1} \right\},$$

Although there are no closed-form solutions for  $\tilde{\delta}(0)$  and  $\tilde{\nu}(0)$ , these equations fit under the framework of a standard inference problem (see Definition 6.2 in (Couillet and Debbah, 2011)). The fixed point iteration algorithm stated in The-

orem 2 is guaranteed to converge to a unique solution  $(\tilde{\delta}(0), \tilde{\nu}(0))$  (see Theorem 6.18 in (Couillet and Debbah, 2011)), denoted  $(\tilde{\delta}, \tilde{\nu})$  in the equations leading up to Theorem 2. So, overall we have

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \bar{\mathbf{Q}} \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \mathbf{A}_0 \} + \frac{1}{n_1} \text{tr} \{ \mathbf{A}_1 \},$$

and

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \asymp -\frac{1}{2} \boldsymbol{\mu}^T \bar{\mathbf{Q}} \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \mathbf{A}_0 \} - \frac{1}{n_1} \text{tr} \{ \mathbf{A}_1 \} \right),$$

where

$$\begin{aligned} \bar{\mathbf{Q}} &:= \lim_{\gamma \rightarrow 0} \bar{\mathbf{Q}}_\gamma \\ &= \left( \frac{n_0 - 1}{n - 2} \frac{1}{1 + \tilde{\delta}(0)} \boldsymbol{\Sigma}_0 + \frac{n_1 - 1}{n - 2} \frac{1}{1 + \tilde{\nu}(0)} \boldsymbol{\Sigma}_1 \right)^{-1}. \end{aligned}$$

### Derivation of $\bar{m}_1$

Similarly, the problem of deriving this deterministic equivalent can be further decomposed into deriving the following additional convergence statements

$$\hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \asymp \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \} - \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \} \right)$$

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \asymp \frac{1}{2} \boldsymbol{\mu}^T \bar{\mathbf{Q}} \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} \text{tr} \{ \mathbf{A}_0 \} - \frac{1}{n_1} \text{tr} \{ \mathbf{A}_1 \} \right)$$

which can be proven in a similar way to the terms composing  $m_0$ .

## Derivation of $\bar{\sigma}_0^2$

The discriminant variance  $\sigma_0^2$  can be expressed as

$$\sigma_0^2 = (1 - \alpha)^2 \rho^2 \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\mu}} + \alpha^2 \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} + 2\alpha(1 - \alpha) \rho \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}.$$

The problem of deriving this deterministic equivalent can be further decomposed into deriving the following additional convergence statements

$$\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \boldsymbol{\Sigma}_0 \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0^2 \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_1 \}$$

$$\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \mathbf{A}_0 \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \mathbf{A}_0 \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_0 \mathbf{A}_1 \}$$

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \tilde{\mathbf{Q}}_0 \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \tilde{\mathbf{Q}}_0 \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \tilde{\mathbf{Q}}_0 \}$$

The first two results can be shown using the same techniques as above. The third result needs special treatment, as it involves a double resolvent. Using the result for double resolvents in (Benaych-Georges and Couillet, 2016), in conjunction with taking  $\gamma \rightarrow 0$ , we have

$$\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \leftrightarrow \tilde{\mathbf{Q}}_0.$$

## Derivation of $\bar{\sigma}_1^2$

Similarly, the problem of deriving this deterministic equivalent can be further decomposed into deriving the following additional convergence statements

$$\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_1 \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \boldsymbol{\Sigma}_1 \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \boldsymbol{\Sigma}_1 \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1^2 \}$$

$$\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_1 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \mathbf{A}_1 \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_1 \mathbf{A}_0 \} + \frac{1}{n_1} \text{tr} \{ \boldsymbol{\Sigma}_1 \mathbf{A}_1 \}$$

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}_1 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \tilde{\mathbf{Q}}_1 \boldsymbol{\mu} + \frac{1}{n_0} \text{tr} \left\{ \boldsymbol{\Sigma}_0 \tilde{\mathbf{Q}}_1 \right\} + \frac{1}{n_1} \text{tr} \left\{ \boldsymbol{\Sigma}_1 \tilde{\mathbf{Q}}_1 \right\}$$

The third convergence statement uses the result

$$\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}_1 \hat{\boldsymbol{\Sigma}}^{-1} \leftrightarrow \tilde{\mathbf{Q}}_1$$

from (Benaych-Georges and Couillet, 2016).

### C.2.2 Common covariances

The following proofs rely heavily on three main facts. Firstly, under the assumption that the two classes have common covariance  $\boldsymbol{\Sigma}$ ,  $\hat{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_i \mathbf{1}}{n_i}$  for some  $\mathbf{Z}_i$  with  $\mathcal{N}(\mathbf{0}, \mathbf{I})$  columns,  $i = 0, 1$ . Secondly,  $\hat{\boldsymbol{\mu}}_i$ ,  $i = 0, 1$  are independent of  $\hat{\boldsymbol{\Sigma}}$ . Finally,  $\hat{\boldsymbol{\Sigma}}$  can be expressed as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \boldsymbol{\Sigma}^{1/2} \bar{\mathbf{Z}} \bar{\mathbf{Z}}^T \boldsymbol{\Sigma}^{1/2}$$

for some  $\mathbf{Z} \in \mathbb{R}^{p \times (n-2)}$  which has i.i.d. entries distributed as  $\mathcal{N}(0, 1)$ . The proofs follow the same line of reasoning as those at the beginning of Section C.2.1.

### Derivation of $\bar{m}_0$

The problem of deriving this deterministic equivalent can be further decomposed into deriving the following convergence statements

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}} &\asymp \boldsymbol{\mu}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \} \\ \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} &\asymp \tau \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right] \\ \hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) &\asymp -\frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} - \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \} \\ \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) &\asymp -\frac{\tau}{2} \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \frac{p}{n_0} + \frac{p}{n_1} \right] \end{aligned}$$

We now derive the second convergence statement in detail. It is mostly representative of the rest of the derivations. The term  $\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$  can be expressed as

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} = \left( \boldsymbol{\mu} + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1} - \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} \right)^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu} + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1} - \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0} \right),$$

where  $\mathbf{Z}_i \in \mathbb{R}^{p \times n_i}$ ,  $i = 0, 1$  has i.i.d.  $\mathcal{N}(0, 1)$  entries. Taking the expectation over  $\mathbf{Z}_i \mathbf{1}$ ,  $i = 0, 1$ , while making use of the fact that  $\hat{\boldsymbol{\Sigma}}$  is independent of  $\hat{\boldsymbol{\mu}}$ , and that  $\frac{\mathbf{Z}_i \mathbf{1}}{n_i} \sim \mathcal{N}\left(\mathbf{0}_p, \frac{1}{n_i} \mathbf{I}_p\right)$ ,  $i = 0, 1$ , we have the following intermediate convergence result

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \boldsymbol{\mu}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \left\{ \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \right\}.$$

We have

$$\begin{aligned} \boldsymbol{\mu}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \left\{ \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \right\} &\asymp \boldsymbol{\mu}^T \left( \frac{1}{n-2} \boldsymbol{\Sigma}^{1/2} \bar{\mathbf{Z}} \bar{\mathbf{Z}}^T \boldsymbol{\Sigma}^{1/2} \right)^{-1} \boldsymbol{\mu} \\ &\quad + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \left\{ \boldsymbol{\Sigma} \left( \frac{1}{n-2} \boldsymbol{\Sigma}^{1/2} \bar{\mathbf{Z}} \bar{\mathbf{Z}}^T \boldsymbol{\Sigma}^{1/2} \right)^{-1} \right\} \\ &= \lim_{\gamma \rightarrow 0} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{Q} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \left\{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{Q} \right\} \right], \end{aligned}$$

where  $\mathbf{Q} = \left( \frac{1}{n-2} \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2} \mathbf{W} \mathbf{W}^T \mathbf{D}_{\boldsymbol{\Sigma}}^{1/2} + \gamma \mathbf{I}_p \right)^{-1}$  and  $\mathbf{W} = \mathbf{V}^T \bar{\mathbf{Z}} \in \mathbb{R}^{p \times n-2}$  also has i.i.d. entries distributed as  $\mathcal{N}(0, 1)$  due to invariance of the Gaussian distribution to orthogonal transformations. Using the results in (Hachem et al., 2013), we have

$$\boldsymbol{\mu}^T \mathbf{V} \mathbf{Q} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \left\{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{Q} \right\} \asymp \boldsymbol{\mu}^T \mathbf{V} \mathbf{T} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \left\{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{T} \right\}$$

where

$$\mathbf{T} = -\frac{1}{\gamma} \left( \mathbf{I}_p + \tilde{\delta} \mathbf{D}_{\boldsymbol{\Sigma}} \right)^{-1}$$

and

$$\begin{aligned}\delta &= \frac{1}{n} \text{tr} \left\{ \mathbf{D}_\Sigma \left( -\gamma \left( \mathbf{I}_p + \tilde{\delta} \mathbf{D}_\Sigma \right) \right)^{-1} \right\} \\ \tilde{\delta} &= -\frac{1}{\gamma(1+\delta)}\end{aligned}$$

The desired DE is

$$\lim_{n,p \rightarrow \infty} \lim_{\gamma \rightarrow 0} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{Q} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_\Sigma \mathbf{Q} \} \right] \quad (\text{C.12})$$

To be able to apply the above asymptotic result to this expression, we first need to justify the interchange of the limits in (C.12). This can be done using the Moore-Osgood theorem in a similar way to that shown in Section C.2.1.

$$\begin{aligned}\lim_{\gamma \rightarrow 0} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{Q} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_\Sigma \mathbf{Q} \} \right] &\asymp \lim_{n,p \rightarrow \infty} \lim_{\gamma \rightarrow 0} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{Q} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_\Sigma \mathbf{Q} \} \right] \\ &= \lim_{\gamma \rightarrow 0} \lim_{n,p \rightarrow \infty} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{Q} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_\Sigma \mathbf{Q} \} \right] \\ &= \lim_{\gamma \rightarrow 0} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{T} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_\Sigma \mathbf{T} \} \right]\end{aligned}$$

By making appropriate substitutions in  $\mathbf{T}$  and taking the limit, it can be shown that

$$\lim_{\gamma \rightarrow 0} \mathbf{T} = \tau \mathbf{D}_\Sigma^{-1}$$

under growth condition (d). So overall we obtain

$$\hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}} \asymp \tau \left[ \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right]$$

## Derivation of $\bar{m}_1$

The problem of deriving this deterministic equivalent can be reduced to deriving the following additional convergence statements

$$\begin{aligned}\hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) &\asymp \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{1}{2} \left( \frac{1}{n_0} - \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \} \\ \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) &\asymp \frac{\tau}{2} \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{p}{n_0} - \frac{p}{n_1} \right]\end{aligned}$$

The proofs are similar to those in Section C.2.2

### Derivation of $\bar{\sigma}_0^2 = \bar{\sigma}_1^2$

The discriminant variance can be expressed as

$$\begin{aligned}\sigma_0^2 = \sigma_1^2 &= \left( \rho \hat{\boldsymbol{\mu}}^T + \alpha \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \right) \boldsymbol{\Sigma} \left( \rho \hat{\boldsymbol{\mu}}^T + \alpha \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \right)^T \\ &= (1 - \alpha)^2 \rho^2 \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\mu}} + \alpha^2 \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} + 2\alpha(1 - \alpha) \rho \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}\end{aligned}$$

The problem of deriving this deterministic equivalent can be reduced to deriving the following additional convergence statements

$$\begin{aligned}\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\mu}} &\asymp \boldsymbol{\mu}^T \boldsymbol{\Sigma} \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma}^2 \} \\ \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} &\asymp \tau \left[ \boldsymbol{\mu}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \} \right] \\ \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} &\asymp \tau^3 \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right]\end{aligned}$$

The last convergence claim involves a double resolvent and therefore we include its derivation here. Using the same technique as before to remove the randomness coming from the sample means, we can show that

$$\begin{aligned}\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} &\asymp \boldsymbol{\mu}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \} \\ &\asymp \lim_{\gamma \rightarrow 0} \left[ \boldsymbol{\mu}^T \mathbf{V} \mathbf{Q} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{Q} \mathbf{V}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{Q} \mathbf{D}_{\boldsymbol{\Sigma}} \mathbf{Q} \} \right]\end{aligned}$$

Using the result in (Kammoun et al., 2019) for double resolvents and by interchanging limits as before, we can show that the double resolvent introduces a

correction factor of  $\frac{1}{1-\frac{p}{n}}$  (in addition to the  $\frac{1}{1-\frac{p}{n}}$  introduced by each of the sample covariance matrices) and thus we have

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \tau^3 \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right]$$

### C.3 Derivation of the G-estimator of the probability of misclassification

Deriving the G-estimators  $\hat{m}_0$ ,  $\hat{m}_1$ ,  $\hat{\sigma}_0^2$ , and  $\hat{\sigma}_1^2$ , reduces to deriving the G-estimators of the constituent quadratic forms of  $m_0$ ,  $m_1$ ,  $\sigma_0^2$ , and  $\sigma_1^2$  that are functions of true statistics. This is the approach taken in what follows.

#### C.3.1 Distinct covariances

##### Derivation of $\hat{m}_0$

Deriving the G-estimator for  $m_0$  decomposes into deriving G-estimators of the following terms

$$\hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$$

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$$

Comparing the DE of the plugin estimator  $\hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\mu}}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$  to that of  $\hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$ , we see that we need to add a correction of  $\frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \}$  to the plugin estimator. It is easy to show that

$$\frac{1}{n_0} \text{tr} \{ \hat{\boldsymbol{\Sigma}}_0 \} \asymp \frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \}$$

from which it follows that

$$\hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\mu}}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \frac{1}{n_0} \text{tr} \{ \hat{\boldsymbol{\Sigma}}_0 \} \asymp \hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$$

By comparing the DE of the plugin estimator  $\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \hat{\boldsymbol{\mu}}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$  to that of  $\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$ , we observe that we must add a correction of  $\frac{1}{n_0} \text{tr} \{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \}$

to the plugin estimator. A G-estimator for  $\frac{1}{n_0} \text{tr} \left\{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \right\}$  is derived as follows.

Expressing  $\hat{\boldsymbol{\Sigma}}_0$  and  $\hat{\boldsymbol{\Sigma}}_1$  as

$$\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n-2} \sum_{i=1}^{n-2} \tilde{\mathbf{y}}_{i0} \tilde{\mathbf{y}}_{i0}^T$$

where  $\tilde{\mathbf{y}}_{i0} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$ ,  $i = 1, 2, \dots, n-2$ , and

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n-2} \sum_{i=1}^{n-2} \tilde{\mathbf{y}}_{i1} \tilde{\mathbf{y}}_{i1}^T$$

where  $\tilde{\mathbf{y}}_{i1} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ ,  $i = 1, 2, \dots, n-2$ , then

$$\begin{aligned} \frac{1}{n_0} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\Sigma}}^{-1} \right\} &= \frac{1}{n-2} \sum_{i=1}^{n-2} \frac{1}{n_0} \text{tr} \left\{ \tilde{\mathbf{y}}_{0i}^T \left( \frac{1}{n-2} \sum_j \tilde{\mathbf{y}}_{0j} \tilde{\mathbf{y}}_{0j}^T + \frac{1}{n-2} \sum_k \tilde{\mathbf{y}}_{1k} \tilde{\mathbf{y}}_{1k}^T \right)^{-1} \tilde{\mathbf{y}}_{0i} \right\} \\ &= \frac{1}{n-2} \sum_{i=1}^{n-2} \frac{\frac{1}{n_0} \tilde{\mathbf{y}}_{0i}^T \mathbf{Q}_i \tilde{\mathbf{y}}_{0i}}{1 + \frac{1}{n-2} \tilde{\mathbf{y}}_{0i}^T \mathbf{Q}_i \tilde{\mathbf{y}}_{0i}} \\ &\asymp \frac{1}{n-2} \sum_{i=1}^{n-2} \frac{\frac{1}{n_0} \text{tr} \{ \mathbf{Q}_i \boldsymbol{\Sigma}_0 \}}{1 + \frac{1}{n-2} \text{tr} \{ \mathbf{Q}_i \boldsymbol{\Sigma}_0 \}} \\ &\asymp \frac{\frac{1}{n_0} \text{tr} \left\{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \right\}}{1 + \frac{1}{n-2} \text{tr} \left\{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \right\}} \end{aligned}$$

where  $\mathbf{Q}_i = \left( \frac{1}{n-2} \sum_{j \neq i} \tilde{\mathbf{y}}_{0j} \tilde{\mathbf{y}}_{0j}^T + \frac{1}{n-2} \sum_k \tilde{\mathbf{y}}_{1k} \tilde{\mathbf{y}}_{1k}^T \right)^{-1}$ . Rearranging, we have

$$\frac{n-2}{n_0} \lambda_0 \asymp \frac{1}{n_0} \text{tr} \left\{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \right\}$$

and so overall,

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \hat{\boldsymbol{\mu}}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \frac{n-2}{n_0} \lambda_0 \asymp \hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right)$$

## Derivation of $\hat{m}_1$

Using the same approach as is used for deriving  $\hat{m}_0$ , it can be shown that

$$\hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\mu}}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) - \frac{1}{n_1} \text{tr} \left\{ \hat{\boldsymbol{\Sigma}}_1 \right\} \asymp \hat{\boldsymbol{\mu}}^T \left( \boldsymbol{\mu}_1 - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right)$$

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \hat{\boldsymbol{\mu}}_1 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) - \frac{n-2}{n_1} \lambda_1 \asymp \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \left( \boldsymbol{\mu}_1 - \frac{\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1}{2} \right)$$

## Derivation of $\hat{\sigma}_0^2$

Deriving the G-estimator for  $\sigma_0^2$  decomposes into deriving G-estimators of  $\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\mu}}$ ,  $\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$ , and  $\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$ . We can easily show

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\mu}}$$

which takes care of the first term. We will now show that

$$(1 + \lambda_0) \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$$

Firstly,

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} &= \frac{1}{n-2} \sum_{i=1}^{n-2} \hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{y}}_{i0} \tilde{\boldsymbol{y}}_{i0}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \\ &= \frac{1}{n-2} \sum_{i=1}^{n-2} \frac{\hat{\boldsymbol{\mu}}^T \tilde{\boldsymbol{y}}_{i0} \tilde{\boldsymbol{y}}_{i0}^T \left( \frac{1}{n-2} \sum_{j \neq i} \tilde{\boldsymbol{y}}_{j0} \tilde{\boldsymbol{y}}_{j0}^T + \frac{1}{n-2} \sum_{k=1}^{n-2} \tilde{\boldsymbol{y}}_{k1} \tilde{\boldsymbol{y}}_{k1}^T \right)^{-1} \hat{\boldsymbol{\mu}}}{1 + \frac{1}{n-2} \tilde{\boldsymbol{y}}_{i0}^T \left( \frac{1}{n-2} \sum_{j \neq i} \tilde{\boldsymbol{y}}_{j0} \tilde{\boldsymbol{y}}_{j0}^T + \frac{1}{n-2} \sum_{k=1}^{n-2} \tilde{\boldsymbol{y}}_{k1} \tilde{\boldsymbol{y}}_{k1}^T \right)^{-1} \tilde{\boldsymbol{y}}_{i0}} \\ &\asymp \frac{\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}}{1 + \frac{1}{n-2} \text{tr} \left\{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \right\}} \end{aligned}$$

which means that

$$\left( 1 + \frac{1}{n-2} \text{tr} \left\{ \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \right\} \right) \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma}_0 \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}$$

The final expression is obtained by substituting the G-estimator of  $\frac{1}{n-2}\text{tr}\left\{\Sigma_0\hat{\Sigma}^{-1}\right\}$  derived previously.

In a similar way, it can be shown that

$$(1 + \lambda_0)^2 \hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \hat{\Sigma}_0 \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \Sigma_0 \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$$

### Derivation of $\hat{\sigma}_1^2$

In a similar manner, we derive the following convergence relations for the constituent terms of  $\sigma_1^2$

$$\hat{\boldsymbol{\mu}}^T \hat{\Sigma}_1 \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \Sigma_1 \hat{\boldsymbol{\mu}}$$

$$(1 + \lambda_1) \hat{\boldsymbol{\mu}}^T \hat{\Sigma}_1 \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \Sigma_1 \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$$

$$(1 + \lambda_1)^2 \hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \hat{\Sigma}_1 \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \Sigma_1 \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$$

### C.3.2 Common covariances

#### Derivation of $\hat{m}_0$

Expressing  $m_0$  as

$$\begin{aligned} m_0 &= \left( \rho \hat{\boldsymbol{\mu}}^T + \alpha \hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \mathbf{P}_{\hat{\boldsymbol{\mu}}} \right) \left( \hat{\boldsymbol{\mu}}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} + \boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0 \right) \\ &= (1 - \alpha) \rho \hat{\boldsymbol{\mu}}^T \left( \hat{\boldsymbol{\mu}}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) + \alpha \hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \left( \hat{\boldsymbol{\mu}}_0 - \frac{\hat{\boldsymbol{\mu}}_0 + \hat{\boldsymbol{\mu}}_1}{2} \right) \\ &\quad + (1 - \alpha) \rho \hat{\boldsymbol{\mu}}^T (\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0) + \alpha \hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} (\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0), \end{aligned}$$

we see that G-estimators for  $\hat{\boldsymbol{\mu}}^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0)$  and  $\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0)$  are needed. By substituting

$$\hat{\boldsymbol{\mu}}_0 = \boldsymbol{\mu}_0 + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_0 \mathbf{1}}{n_0},$$

$$\hat{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \frac{\boldsymbol{\Sigma}^{1/2} \mathbf{Z}_1 \mathbf{1}}{n_1},$$

and taking the expectation over  $\mathbf{Z}_0 \mathbf{1}$  and  $\mathbf{Z}_1 \mathbf{1}$  in  $\hat{\boldsymbol{\mu}}^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0)$ , we obtain

$$\hat{\boldsymbol{\mu}}^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0) \asymp \frac{1}{n_0} \text{tr}\{\boldsymbol{\Sigma}\}.$$

We can easily show that

$$\frac{1}{n_0} \text{tr}\{\hat{\boldsymbol{\Sigma}}\} \asymp \frac{1}{n_0} \text{tr}\{\boldsymbol{\Sigma}\}$$

by substituting  $\frac{1}{n-2} \boldsymbol{\Sigma}^{1/2} \bar{\mathbf{Z}} \bar{\mathbf{Z}}^T \boldsymbol{\Sigma}^{1/2}$  for  $\hat{\boldsymbol{\Sigma}}$  and taking the expectation. Thus, we have

$$\frac{1}{n_0} \text{tr}\{\hat{\boldsymbol{\Sigma}}\} \asymp \hat{\boldsymbol{\mu}}^T(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0).$$

Through a similar derivation, we obtain

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\mu}_0 - \hat{\boldsymbol{\mu}}_0) \asymp \frac{1}{n_0} \text{tr}\{\boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1}\}.$$

To find the G-estimator of this quantity, replace  $\boldsymbol{\Sigma}$  with its estimate and then express this as a function of the original quantity as follows. First express  $\hat{\boldsymbol{\Sigma}}$  as

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-2} \sum_{i=1}^{n-2} \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T,$$

where  $\tilde{\mathbf{y}}_i \sim \mathcal{N}(\mathbf{0}, \Sigma)$ ,  $i = 1, 2, \dots, n-2$ . So we have

$$\begin{aligned} \frac{1}{n_0} \text{tr} \left\{ \Sigma \hat{\Sigma}^{-1} \right\} &= \frac{1}{n-2} \sum_{i=1}^{n-2} \frac{1}{n_0} \text{tr} \left\{ \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T \left( \frac{1}{n-2} \sum_{j=1}^{n-2} \tilde{\mathbf{y}}_j \tilde{\mathbf{y}}_j^T \right)^{-1} \right\} \\ &= \frac{1}{n-2} \sum_{i=1}^{n-2} \frac{1}{n_0} \text{tr} \left\{ \tilde{\mathbf{y}}_i^T \left( \frac{1}{n-2} \sum_{j=1}^{n-2} \tilde{\mathbf{y}}_j \tilde{\mathbf{y}}_j^T \right)^{-1} \tilde{\mathbf{y}}_i \right\} \\ &= \frac{1}{n-2} \sum_{i=1}^{n-2} \frac{1}{n_0} \text{tr} \left\{ \tilde{\mathbf{y}}_i^T \left( \frac{1}{n-2} \sum_{j \neq i} \tilde{\mathbf{y}}_j \tilde{\mathbf{y}}_j^T + \tilde{\mathbf{y}}_i \tilde{\mathbf{y}}_i^T \right)^{-1} \tilde{\mathbf{y}}_i \right\} \\ &= \frac{\frac{1}{n_0} \tilde{\mathbf{y}}_i^T \mathbf{Q}_i \tilde{\mathbf{y}}_i}{1 + \frac{1}{n-2} \tilde{\mathbf{y}}_i^T \mathbf{Q}_i \tilde{\mathbf{y}}_i}, \end{aligned}$$

where  $\mathbf{Q}_i = \left( \frac{1}{n-2} \sum_{j \neq i} \tilde{\mathbf{y}}_j \tilde{\mathbf{y}}_j^T \right)^{-1}$  and the last line follows from applying the matrix inversion lemma in (Müller and Debbah, 2016). It can be shown that

$$\frac{\frac{1}{n_0} \tilde{\mathbf{y}}_i^T \mathbf{Q}_i \tilde{\mathbf{y}}_i}{1 + \frac{1}{n-2} \tilde{\mathbf{y}}_i^T \mathbf{Q}_i \tilde{\mathbf{y}}_i} \asymp \frac{\frac{1}{n_0} \text{tr} \{ \Sigma \hat{\Sigma}^{-1} \}}{1 + \frac{1}{n-2} \text{tr} \{ \Sigma \hat{\Sigma}^{-1} \}}.$$

Therefore,

$$\frac{1}{n_0} \text{tr} \{ \hat{\Sigma} \hat{\Sigma}^{-1} \} = \frac{p}{n_0} \asymp \frac{\frac{1}{n_0} \text{tr} \{ \Sigma \hat{\Sigma}^{-1} \}}{1 + \frac{1}{n-2} \text{tr} \{ \Sigma \hat{\Sigma}^{-1} \}}$$

and solving for the original quantity, we have

$$\frac{\frac{p}{n_0}}{1 - \frac{p}{n-2}} \asymp \frac{1}{n_0} \text{tr} \left\{ \Sigma \hat{\Sigma}^{-1} \right\}.$$

## Derivation of $\hat{m}_1$

The G-estimators for  $\hat{\boldsymbol{\mu}}^T (\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)$  and  $\hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}_1)$  are derived in a similar fashion to the previous section.

## Derivation of $\hat{\sigma}_0^2 = \hat{\sigma}_1^2$

We need G-estimators for the terms  $\hat{\boldsymbol{\mu}}^T \Sigma \hat{\boldsymbol{\mu}}$ ,  $\hat{\boldsymbol{\mu}}^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$ , and  $\hat{\boldsymbol{\mu}}^T \Sigma \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}$ . It can be easily shown that

$$\hat{\boldsymbol{\mu}}^T \hat{\Sigma} \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \Sigma \hat{\boldsymbol{\mu}}.$$

From Appendix C.2.2, we have

$$\hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \tau^3 \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right]$$

If we replace  $\boldsymbol{\Sigma}$  by its estimator, we have

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \\ &\asymp \tau \left[ \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \frac{p}{n_0} + \frac{p}{n_1} \right] \end{aligned}$$

therefore,

$$\tau^2 \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}.$$

From Appendix C.2.2, we know that

$$\hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} \asymp \tau \left[ \boldsymbol{\mu}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \} \right].$$

If we replace  $\boldsymbol{\Sigma}$  by its estimator, we have

$$\begin{aligned} \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}} \\ &\asymp \boldsymbol{\mu}^T \boldsymbol{\mu} + \left( \frac{1}{n_0} + \frac{1}{n_1} \right) \text{tr} \{ \boldsymbol{\Sigma} \}, \end{aligned}$$

where the last line is from Appendix C.2.2. Therefore,

$$\tau \hat{\boldsymbol{\mu}}^T \hat{\boldsymbol{\mu}} \asymp \hat{\boldsymbol{\mu}}^T \boldsymbol{\Sigma} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}.$$

## Appendix D

### Papers Accepted, Submitted, and Under Preparation

- Lama B. Niyazi, Abla Kammoun, Hayssam Dahrouj, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri, “Asymptotic Analysis of an Ensemble of Randomly Projected Linear Discriminants”, *Accepted to the IEEE Journal on Selected Areas in Information Theory*, November 2020.
- Lama B. Niyazi, Abla Kammoun, Hayssam Dahrouj, Mohamed-Slim Alouini, and Tareq Y. Al-Naffouri, “Weight Vector Tuning and Asymptotic Analysis of Binary Linear Classifiers”, *Accepted to the IEEE Open Journal of Signal Processing*, July 2022.
- Lama B. Niyazi, Abla Kammoun, Hayssam Dahrouj, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini, “An Asymptotic Study of Discriminant and Vote-Averaging Schemes for Randomly-Projected Linear Discriminants”, *Submitted to the Journal of Machine Learning Research*, November 2022.
- Lama B. Niyazi, Abla Kammoun, Hayssam Dahrouj, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini, “General Shrinkage for Linear Discriminant Analysis Classification”, *Under preparation*.