Australian National University

Benign Overfitting in High-Dimensional Linear Discriminant Analysis

Kehan Zhao

November 2023

Supervised by Dr Yanrong Yang and Prof Hanlin Shang

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Statistics with Honours in Statistics at the Australian National University.

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University, and, to the best of my knowledge and belief, contains no material published or written by another person, except where due reference is made in the thesis.

Kehan Zhao

For my grandparents.

Acknowledgements

Firstly, I would like to thank my supervisor, Dr. Yanrong Yang, for accepting me as her Honours student and tirelessly supporting me throughout this journey. Not only has she guided me through the process of doing research in Statistics, but she has also taught me to be resilient and to develop skills that could benefit me later in life. Secondly, I would like to thank my associate supervisor, Professor Hanlin Shang. His advice is always direct, practical and assuring. This project would not have been possible without their help. I would also like to thank Dr. Bronwyn Loong, for convening this program and creating a supportive environment for all the Honours Students. Jonathan Tammen, my mentor from Australian Signals Directory (ASD), has also been incredibly helpful throughout this journey.

I would like to extend my gratitude to the Research School of Finance, Actuarial Studies and Applied Statistics (RSFAS) and Australian Signals Directory (ASD) for their financial support.

I would also like to thank my family and friends for their consistent encouragement and unwavering belief. Finally, I want to congratulate myself for not giving up.

Abstract

The Double Descent phenomenom observed in deep neural network has called into question standard machine learning practices. Over-parmeterization, or increasing features p relative to sample size n, may have the effect of decreasing prediction error beyond the classical U-shaped error. However, little research has related this phenomenon to Linear Discriminant Analysis (LDA), a long-standing classification method that finds the linear combination of features that best differentiates between two classes.

In this paper, we leverage recent Random Matrix results to derive the convergence of LDA's error rate in the under-parameterized regime. Particularly, we are interested in how the error varies with the asymptotic ratio of dimension and sample size: $p/n \to \gamma, \gamma \in (0, 1)$. Comparing with previous work, we use a less restrictive assumption where data is not assumed to be normally distributed. Our theoretical result shed light on how LDA's model flexibility influences its error rate in when p and n are both large.

Simulation is conducted to demonstrate the shape of the error curve for LDA with Moore-Penrose pseudo-inverse under various controlled data settings, with the focus on both under-parameterized and over-parameterized region. We notice the occurrence of Double Descent and discover that global minimum error almost always occurs in the over-parameterized regime. Finally, we substantiate the findings in a real data analysis using cancer classification dataset and provide practical implications for practitioners working with LDA in high-dimensional settings. The results underscore the benefit of overparamterization and further contribute the understanding of model risk in the modern data-centric era.

Contents

Notation and terminology

Notation

p	Dimension
n	Sample Size
γ	Limiting ratio of dimension and sample size $\frac{p}{n} \to \gamma$, as $n \to \infty$, $p \to \infty$
Σ	Population Covariance Matrix
S	Sample Covariance Matrix
μ	Population Mean Vector
$\bar{X}$	Sample Mean Vector
Φ	Culmulative Normal Distribution
δ	Difference between the population mean $\mu_1 - \mu_2$
R_{LDA}	Error Rate of Linear Discriminant Analysis
$\xrightarrow{a.s.}$	Converges almost surely

Terminology

LDA	Linear Discriminant Analysis
PLDA	Pseudo-inverse Linear Discriminant Analysis
RMT	Random Matrix Theory
ESD	Empirical Spectral Distribution
LSD	Limiting Spectral Distribution
M-P Law	Marchenko-Pastur Law

1 Introduction

As declared by mathematician Clive Humby in 2006 "Data is the new oil", data has become the valuable asset that fuels the digital economy, steers decision-making, and connects global systems. From digital sensors, social media platforms, and online records to the Internet of Things devices, the ubiquitous generation and consumption of data have up-scaled its potential for knowledge extraction. However, this transformation presents challenges: the complex, unprocessed data requires advanced analytical tools to turn them into digestible information. Out of these tools, machine learning emerges as a powerful way to discern patterns and make predictions. Extending upon statistical models, it develops algorithms through learning from a set of training data.

In traditional statistics, we expect our observations, or number of experimental units n to be large whilst only dealing with a small number of features p. Familiar theoretical results, such as Laws of Large Numbers and Central Limit Theorem rely on the setting of n tending to infinity and p remaining small. However, due to the advancement in data collection avenues and computing power, the data collected are increasingly becoming high-dimensional, meaning that it has more features relative to the number of observations $p > n$ (Johnstone and Titterington, 2009). For example, high dimensionality is prominent in bio-informatics, where microarray technology enables the probes of an entire genome to be placed on a chip. Xing et al. (2001) worked with microarray data of 72 data points in a 7130-dimensional space and applied feature selection to improve classification results. In facial recognition, the number of pixels in an image is often much larger than the number of images, which requires modification of classification methods (Yu and Yang, 2001). Many methods, such as ordinary least squares regression and Multiple Signal Classification, fail under such setting. Thus, there has been a

spur in statistical research focusing on applying methods in high dimensions and understanding the performance of models in such setting. Notably, a surprising phenomenon has been observed: as the number of predictors p increase beyond sample size n, a models' prediction error may also decrease, which is also referred to as "benign overfitting" or "double descent" (Belkin et al., 2019; Hastie et al., 2022).

In the succeeding chapters, we will navigate the intricate relationship between model flexibility and prediction accuracy in the context of Linear Discriminant Analysis (LDA) using the tools of Random Matrix Theory (RMT). RMT deals with matrices with random variables as entries and allows for analysis in the high dimensional asymptotic setting where both p and n tend to infinity. We will explore LDA in such setting through theory, simulation and data analysis to investigate if this model exhibits benign overfitting. By doing so, we will fill a gap in the literature that understands the performance of machine learning models in high-dimensions.

1.1 Background: Linear Discriminant Analysis

1.1.1 Supervised Machine Learning

In supervised machine learning, a set of labelled training data consisting of input and output is used to train the model. For classification problems, the outcome is discrete or belongs to categorical classes.

Suppose we observe n training samples $(X_i, Y_i) \in \mathbb{R}^p$ drawn independently from an unknown distribution $\mathcal{D}$. For each input data X_i, there is an associated response measurement, Y_i.

$$(X_1, Y_1), ...(X_n, Y_n)$$

The predictor X_i can be multivariate, meaning that each observation contains multiple attributes, which can also be referred to as the dimension p. The goal is then to find a function which relates the predictor to the response in order to predict the membership of unseen data where X is observed but the associated Y is not. The training process typically involves adjust-

ing internal parameter values to minimize a loss function ↕. The performance of a predictor can be evaluated by using n_t number of out-of-sample data points, which is not part of the training set $\boldsymbol{X}_t := [X_1, X_2, \ldots, X_{n_t}]$. We can use the misclassification error, or the proportion of new observations that are incorrectly labelled out of all testing observations (n_t):

$$\frac{1}{n_t}\sum_{i=1}^{n_t} I(Y_i \neq \widehat{Y}_i), \qquad i = 1 \ldots, n_t$$

where $\widehat{Y}_i$ is the predicted class label.Many methods have been proposed for this problem, including linear discriminant analysis, logistic regression, support vector machine, and nearest neighbour classifier.

1.1.2 Definitions

Before introducing the LDA model, we will first introduce a few key definitions.

Definition 1.1 (Mean Vector)**.** *Given a data matrix $(n \times p)$ $\boldsymbol{X}$ consisting of n random multivariate observations $(X_1, X_2, \ldots, X_n)^\top$ where each observation X_i is a p dimensional vector, the mean vector μ can be written as*

$$\mu = \mathbb{E}[\boldsymbol{X}]$$

and the sample mean vector is a $(p \times 1)$ column vector containing the average value of the n observations for each of the p variables:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Definition 1.2 (Covariance Matrix)**.** *Consider the same data matrix $\boldsymbol{X}$ as above with mean vector μ. Its covariance matrix is defined as*

$$\Sigma = Var(\boldsymbol{X}) = \mathbb{E}[(\boldsymbol{X} - \mu)(\boldsymbol{X} - \mu)^\top]$$

The population covariance matrix Σ is symmetric and positive semi-definite matrix with main

diagonal elements containing variances of each variable and off-diagonal elements containing covariance between variables. The $(p \times p)$ *centered sample covariance matrix of* $\boldsymbol{X}$*, denoted as* S *is then defined as:*

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top$$

if the data has mean $\bar{X} = 0$*, the sample covariance matrix becomes:*

$$S := \frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^\top$$

If the columns of $\mathbf{X}$ are independently distributed as $N(0, \Sigma)$ for some positive semi-definite matrix Σ, the distribution of $(n-1)S$ follows a p-dimensional Wishart distribution $W_p(n-1, \Sigma)$, where $n-1$ stands for the degrees of freedom and Σ is the scale parameter.

1.1.3 Linear Discriminant Analysis as a Classifier

One of the popular classification methods is linear discriminant analysis (LDA). LDA was first introduced for two-class classification problem by Sir Ronald Fisher in 1936 (Fisher, 1936) and later extended to multi-class classification. The method assumes that each p dimensional input $X \in \mathbb{R}^p$ belong to one of the k groups G_k, denoted by the response $Y \in \{1, \ldots, K\}$. X follows a multivariate normal distribution within each class, with different mean vector μ_k and a common covariance matrix Σ. We shall focus on binary classification, the case where there are two response classes ($K = 2$).

$$X|(Y = k) \sim N(\mu_k, \Sigma) \quad (k = 1, 2) \tag{1.1}$$

For a given input, LDA uses the classification strategy that assigns it to the response class with the highest $f_k(X)$, which is the class-conditional density of X belonging in group G_k. This is called the Bayes classification rule (Hastie et al., 2009). The conditional density function is then the probability density function of the multivariate normal distribution with appropriate

group parameters.

$$f_k(X) = \mathbb{P}(X|Y = k) \tag{1.2}$$

$$f_k(X) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp(-\frac{(X-\mu_k)^\top \Sigma^{-1}(X-\mu_k)}{2}) \tag{1.3}$$

Using the Bayes rule, we find highest posterior probability of an observation belonging to a class, given the predictor X value.

$$\text{Bayes Rule:} \quad \mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)} \tag{1.4}$$

$$\mathbb{P}(Y = k|X) = \frac{\pi_k f_k(X)}{\sum_{k=1}^K \pi_k f_k(X)} \tag{1.5}$$

Note that π_k is the prior probability that a randomly chosen observation comes from the k^{th} class and $\sum_{k=1}^K \pi_k = 1$. (We will assume π_k to be $1/2$ in this analysis).

Since we are assuming equal covariance matrix across group, we can simply the expression by taking the natural logarithm and look-at the log ratio. We arrive at the decision rule where X is classified to the k^{th} class with the highest $d_k(X)$, or the discriminant function value.

$$\widehat{Y} = \underset{k=1,\ldots,K}{\text{argmax}} \mathbb{P}(Y = k|X) = \underset{k=1,\ldots,K}{\text{argmax}} \{d_k(X)\} \tag{1.6}$$

$$d_k(X) = X^\top \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^\top \Sigma^{-1}\mu_k + \log \pi_k \tag{1.7}$$

It is obvious that $d_k(X)$ is a linear function of X, it leads to a linear decision boundary between the classes. Since we assumed the distribution of data, estimating the population parameters μ_k, Σ, π_k is necessary. Often, the maximum likelihood method is used using n training data,

where n_k is the number of observations in group k and $\sum_{k=1}^{K} n_k = n$.

$$\widehat{\pi}_k = \frac{n_k}{n}$$
$$\widehat{\mu}_k = \bar{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki}$$
$$\widehat{\Sigma} = S = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^\top$$

Due to LDA's simplicity and optimality (Anderson, 2003), it remains popular in practice and has competitive performance for many real-world data sets (Cai and Liu, 2011; Fan et al., 2012). There are many extensions to LDA including Quadratic Discriminant Analysis (QDA), where each class is assumed to have a different population covariance matrix, making the decision boundary non-linear. Regularized Discriminant Analysis (RLDA) (Friedman, 1989) where the sample covariance matrix is regularized by a tuning parameter λ. Mixture Discriminant Analysis (MDA) (Hastie and Tibshirani, 1996) also allows for classifying classes that are mixtures of several Gaussian distributions by estimating the parameters using Expectation-Maximization algorithm.

As a classification tool, LDA has been applied in micro-array data, satellite image classification, face recognition, document classification, and speech classification (Michie et al., 1994). LDA can also act as a dimension-reduction tool at the pre-processing step for other algorithms. The new axes found by the discriminant functions reduce the original explanatory variables to $K - 1$. In a way, LDA is similar to Principal Component Analysis (PCA), which focuses on conducting dimension reduction through maximizing data variance, but it also uses class labels to find the maximal separation between classes. We will now consider LDA's performance in the high dimensional setting.

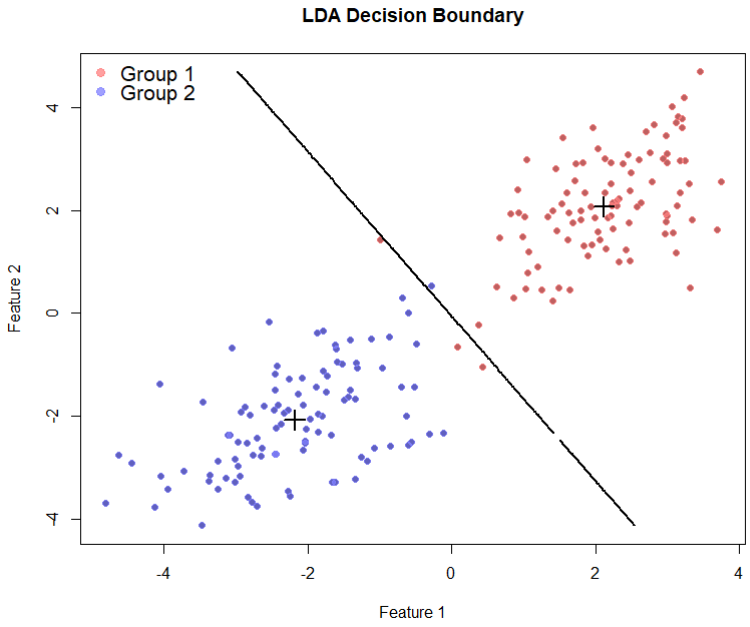


Figure 1.1: Visualization of LDA binary classification linear decision boundary LDA ($n_1 = n_2 = 100, p = 2, \mu_1 = (2,2)^\top, \mu_2 = (-2,-2)^\top, \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$)

1.1.4 LDA in High Dimensional Setting

Moore-Penrose Pseudo-inverse Approach

In the high dimensional case for binary classification, when the number of dimension p is greater than two minus the sample size $n-2$, the sample covariance matrix is not invertible due to rank deficiency. One way of solving this problem is using the generalized Moore-Penrose pseudo-inverse, which always exists for non-full rank matrices. The Moore-Penrose pseudo-inverse exists for any matrix (full rank or non-full rank) and is unique. It was first proposed by Moore (1920) and rediscovered by Penrose (1955). It is useful for dealing with least square problems, such as finding an optimal solution for linear equation $Ax = y$, where A is an $m \times n$ matrix.

Definition 1.3. *Let $A \in \mathbb{R}^{m \times n}$, A^+ is the Moore-Penrose pseudoinverse of A, if and only if it satisfies the following conditions:*

1) $AA^+A = A$

2) $A^+AA^+ = A$

3) $(AA^+)^\top = AA^+$

4) $(A^+A)^\top = A^+A$

The pseudo-inverse is obtained by finding the subspace spanned by the eigenvectors corresponding to non-zero eigenvalues of A. It is also the least square solution to the condition $AA^+ = I$. Computationally, the pseudo-inverse of a symmetric matrix can be derived using singular value decomposition (SVD).

$$A = U^\top DU \qquad A^+ = UD^+U^\top$$

U is orthogonal and D is a diagonal matrix with elements as singular values, which are generalizations of eigenvalues. D^+ is found by taking the reciprocal of all non-zero elements and leaving the zeros unchanged, then modifying the shape of the matrix.

Hoyle (2010) outlined the accuracy of the pseudo-inverse covariance matrix in learning problems. The reconstruction error of the pseudo-inverse is quantified using the Frobenius norm of the difference with the true inverse. In the $p < n$ case, the reconstruction error decreases to a minimum around $n/p \simeq 0.4$ and then rises sharply. Hoyle found that the error is dominated by the smallest nonzero eigenvalues of $\widehat{\Sigma}$, which peaks as n approaches p due to the nonzero eigenvalues approaching zero. In the $p > n$ case, it is found that the reconstruction error decreases following $\frac{1}{(1-\frac{n}{p})^3}$ for a wide range of population covariance matrix structure. Therefore, we would expect the accuracy of the classifier constructed using the pseudo-inverse to follow the same pattern.

Review of LDA methods in High Dimensions

It has been shown that the performance of LDA in high dimension is far from optimal Pillo (1976). The pseudo-inverse method can be unstable when there are small observation and large dimensions also make direct matrix operations cumbersome Guo et al. (2007). Various literature have focused on allowing discriminant analysis to work in small sample sizes, and

high-dimensional problems and improving its performance. One approach involves feature selection: applying Principal Component Analysis as a pre-processing step so that the dimensional reduces to less than the sample size (Fukunaga, 2013). However, it is possible that the PCA step may discard dimension containing important discriminate information.

Other works have focused on modifying the estimation of population parameters (Σ, μ_k) in the high dimensional setting. Friedman (1989)'s regularized discriminant analysis solves the invertibility problem by shrink thing sample covariance matrix towards multiple of the identity matrix, which decreases larger eigenvalues and increases smaller ones, but the tuning parameter needs to be selected. Similarly, Qiao et al. (2008) analysed LDA in the high dimensional low sample size setting, and found poor performance due to "data pilling problem". By assuming that there are a large number of redundant variables, the authors proved that introducing variable selection and sparsity in the discriminant vector can eliminate this issue. Dudoit et al. (2002) proposed using only the diagonal element of the sample covariance matrix and setting the off-diagonal elements to 0 before inverting it. This method assumes independence between features and has been proven to have surprisingly good performance. Shao et al. (2011) proposed sparse linear discriminant analysis where the sparseness criterion is used to perform classification and feature selection simultaneously. Cai and Liu (2011) proposed a method of estimating the precision matrix and difference between the mean vectors together to improve performance. Sifaou et al. (2020) assumed a spiked covariance model (covariance matrix is isotropic except for a finite number of directions) and improved the classifier by correcting the bias in the intercept and minimizing the total misclassification rate.

Other methods include shrunken centroids LDA (Tibshirani et al., 2003; Guo et al., 2007) which regularizes Σ and shrinks $\overline{X}$; Modified LDA (MLDA) where the covariance matrix is constructed from the result of an optimization problem (Ledoit and Wolf, 2004); New LDA which modifies the sample covariance matrix by expanding smaller eigenvalues (Thomaz and Gillies, 2005). The classification accuracy of the various methods has also been summarized and compared by Sharma and Paliwal (2015).

1.2 Background: Benign Overfitting

1.2.1 Bias and Variance Trade-off

An important concept in classical statistical learning is the bias-variance trade-off. A model's generalization error, or testing error, is the summation of bias and variance and irreducible error Hastie et al. (2009). Bias refers to the error from modelling a complex true model with an overly simplified model. Bias is high when the underlying structure of the training dataset is not captured by the succinct model and the relationship between the predictor and the target is missed. On the other hand, variance refers to the error arising from the model's sensitivity to the training data. When the model is too complex, noise, such as random fluctuations, is also captured, causing the prediction to be sensitive to small changes in training data. Thus, the model achieves low training error but does not generalize well. Therefore, there is a trade-off between these two types of errors, leading to the commonly observed U-shaped curve when we plot error against a measure of model complexity. The complexity of the model can be determined by factors such as the training time, number of features, and size of the model. The best practice is then to balance under-fitting (high bias) and over-fitting (high variance) to find a sweet point where the model is both simple and complex, minimizing the testing error.

A related concept that causes the trade-off is the curse of dimensionality. Adding more features to the data tends to decrease bias and increase variance. As data becomes sparse relative to the volume of the feature space, the amount of data needs to increase faster to allow the algorithm to uncover meaningful relationships. The intractability of searching through a high-dimensional space and accurately approximating a high-dimensional function leads to increase in error and the curse of dimensionality. Therefore, in order to obtain good performance, many techniques are used to reduce dimension. However, recent observations in machine learning have challenged the notion of this trade-off by observing "blessings of dimensionality" (Chen et al., 2013), or improvement in error when overfitting with large number of features.

1.2.2 Double Descent Phenomenon

In empirical practices of modern machine learning, it is observed that the model's behavior in the over-parameterized region is at odds with the classical understanding of the bias and variance trade-off. When models such as neural networks are trained to exactly fit the data-set and achieve 0 training error, the prediction is observed to have high accuracy on new data beyond a certain point Zhang et al. (2016). This leads to another descent in the testing error beyond the interpolation threshold where the training error is 0.

The phenomenon has spurred intense research interest. In the 2019 paper, *Reconciling modern machine learning practice and the bias-variance trade-off*, Belkin et al. (2019) first coined the term "Double Descent" to describe the error curve shape (Figure 1.2). Through analysing neural network, decision tree, and ensemble methods, it was shown that the test error exhibited a second descent beyond interpolation boundary. Belkin et al. (2019) proposed that the phenomenon exists because of inductive bias. As the number of features tends to infinity, a larger function class is considered which may contain one that has a small norm. The solution approaches the minimum functional norm in the reproducing kernel Hilbert space for Random Fourier Features. Empirically, Nakkiran et al. (2021) investigated the test error in neural network through varying model size, training epoch, and training sample size. The authors draw the conclusions that bigger models and more data can hurt performance and training longer can benefit test error after interpolation boundary.

Statistical work in linear regression helps to provide intuition and theoretical explanation for the phenomenon observed in computer science. Hastie et al. (2022) explored the performance of the minimum norm least square estimator in linear regression. Utilizing random matrix theory, the authors derived the expression of prediction risk in the double asymptotic setting $p \to \infty, n \to \infty, \frac{p}{n} \to \gamma$. The double descent is observed: the risk diverges as $\gamma \to 1$, before descending and achieving a global minimum at $\gamma > 1$. The intuition provided by the authors is that as p grows, the minimum ℓ_2 norm solution to the linear system $Xb = y$ decreases: with more columns of X in the feature matrix, the components of b can be distributed over more

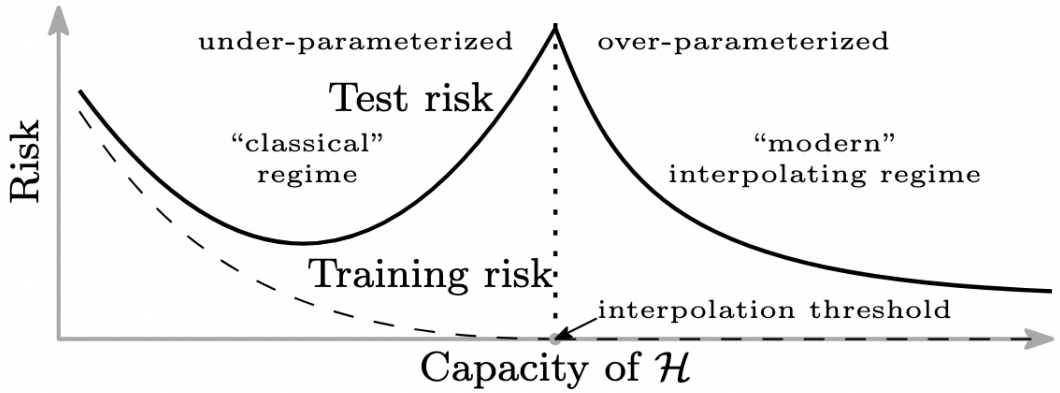


Figure 1.2: Double Descent phenomenon in deep learning models (Belkin et al., 2019)

columns and hence a smaller norm is achieved. In a misspecified model where not all covariates are observed, the double descent phenomenon is more pronounced as bias also decreases with γ. This can be simply explained by the fact that adding features improves approximation capacity by accounting for parts of the true regression function. Similarly, in the latent space model where covariates and coefficient vectors lie in a low dimensional subspace, a monotone decrease in risk as $\gamma \to \infty$ is also observed. The explanation is that each new feature provide more information of the low-dimensional latent variable.

Similarly, Bartlett et al. (2020) also considered a perfect fit of the training data using linear regression with minimum norm solution. By deriving the upper and lower risk bounds, the authors concluded that the existence of many low-variance, unimportant directions in the parameter space is essential in achieving benign overfitting. In the infinite-dimensional setting (separable Hilbert Space), benign overfitting occurs only for a narrow range of conditions where eigenvalues of population covariance matrix decay slowly. In the finite dimension space where p grows faster than n, it occurs with a wider conditions. For benign overfitting to occur, the number of non-zero eigenvalues should be large compared to n, and have a small sum compared to n. The authors also suggested that truncating to a finite-dimensional space is good for model performance in the overfitting regime in deep neural networks.

Double Descent in Classification Models

One of the first demonstrations of double descent was in fact the use of Fisher linear discriminant analysis with pseudo-inverse by Duin (2000). By using simulation and varying n for different values of p, Duin demonstrated that it is possible to construct good classifiers in almost empty spaces. There is a peak in generalization error when $p \simeq n$ and a reduction in error as the dimension keeps increasing. However, Duin only outlined the simulated generalization error curve and did not derive the mathematical expression or the mechanisms behind the double descent for LDA with pseudo-inverse. An empirical example of double descent for image classification is observed by Chen et al. (2013): a single-type Local Binary Pattern (LBP) descriptor can achieve state-of-the-art results in facial recognition accuracy when features are scaled to $100,000$. Theoretical works on double descent for classifiers are less prevalent than regression methods, due to the lack of closed closed-form solution for the error. Montanari et al. (2019) focused on the maximum margin linear classifier and confirmed that under certain assumptions of the decay rate of covariance matrix Σ eigenvalues, the test error is monotonically decreasing in the over-parameterized region. The global minimum achieved at large over-parametrization. Chatterji and Long (2021) worked under the finite sample (fixed n, growing p setting) and proved the error bounds of the maximum margin linear classifier. The error bound approaches Bayes risk as the features go to infinity despite the presence of noise. Deng et al. (2021) studied the misclassification error of gradient descent on logistic loss for logistic and gaussian mixture models at different over-parameterization ratios. Similar to regression, double descent is found under different signal-to-noise ratio settings and the occurrence of global minimum in test error depends on the nature of the training data. However, double descent is not universal in every model. Buschjäger and Morik (2021) challenged the previous view and showed that random forests exhibit single descent rather than double. Nakkiran et al. (2021) also found that using regularization can mitigate double descent and achieve monotonic test error for linear regression and neural networks.

1.3 Thesis Motivation and Structure

The double descent phenomenon has called into question the best practices for selecting the number of features for high-dimensional data. In certain models, dimension reduction might be unnecessary as large numbers of features can achieve a testing error just as low or lower than that of the first minimum. However, there is a gap in our knowledge as no literature has formally linked the concept of benign overfitting to the behavior of LDA's error rate in high-dimensions. The current research has mainly focused on models such as deep neural networks and providing a theoretical explanation for them through deriving asymptotic risk from least square regression. The previous literature on the error expression of LDA has mainly focused on finding the expected misclassification error or its asymptotic expansions in the classical $n > p$ setting. In high dimensional analysis, regularized LDA is often the interest rather than pseudo-inverse LDA.

Given the popularity of LDA and its importance in many applications, we wish to understand the effect of model flexibility on its prediction accuracy and analyze the benign fitting phenomenon through theory, simulation, and data analysis.

Theoretical Analysis: In the theoretical section, we wish to characterize the asymptotic shape of the risk curve for LDA. We will follow the double asymptotic setting (Deev, 1970): the number of dimensions p and total sample size n go to infinity, but their limiting ratio γ remains constant. Due to time constraint, we will focus on deriving the asymptotic convergence of the error in the under-parameterized region $\gamma \in (0, 1)$:

$$p \to \infty$$

$$n \to \infty$$

$$p/n \to \gamma \in (0, 1)$$

By deriving the theoretical error expression, we could answer whether there is a theoretical backing for the error peaking phenomenon of LDA error and better understand how flexibility

can influence performance. The double asymptotic approach is superior to the fixed p setting as it addresses the increasing ubiquity of high-dimensional data and allows for the case where p and n are proportionately large. The setting can also be used to approximate behaviour at large but finite values of p in real applications. We will also provide suggestions on the theoretical error convergence in the over-parameterized regime where Moore-Penrose pseudo-inverse is used.

Simulation: In the simulation section, we will verify the existence of double descent through synthetically generated data. We will observe the shape of the error rate curve for finite values of $\gamma \in (0, a)$ using the pseudo-inverse. We will aim to answer 5 specific research questions by varying the simulation settings to approximate real-life data: effect of sample size, effect of covariance matrix structure variations, normality assumption violations, noisy observations and redundant features.

Real Data Analysis: We will also conduct a real data analysis of the pseudo-LDA model using ARCENE cancer classification dataset (Guyon et al., 2008). The goal is to verify the existence of double descent in a real-life setting. Since the advantage of double descent is to achieve a lower error rate in a high-dimensional setting, we will compare the performance of the model to dimension reduction and other methods of high-dimensional LDA.

Novelty of Result

An early work by Wang and Jiang (2018) derived the theoretical convergence of LDA in the double asymptotic setting in the under-parameterized regime $\gamma < 1$. The concurrent work by Cheng et al. (2022) theoretically explores the relationship between LDA's error rate and model flexibility under label shift (different class proportions). In the over-parameterize regime, the convergence of error expression when pseudo-inverse is used is also derived. However, both of these work uses the assumption that the data is Gaussian ($X_{ki} \overset{iid}{\sim} N(\mu_k, \Sigma)$). The proof technique then depends on the moments of Wishart distribution. In our derivation, we will provide a more general result where the data need not be normally distributed and only has moment restraints. Furthermore, both works have different thematic emphases: Wang and Jiang

(2018) focuses more on correcting bias for high dimensional regularized discriminant analysis and Cheng et al. (2022) focuses on the effects of unbalanced training classes. Our work will tie more to LDA's model flexibility and the application of the double descent phenomenon. Our simulation work also extends upon the error curves illustrated by Duin (2000) to demonstrate a variety of settings. From previous literature and theoretical understanding of the behaviour of pseudo-inverse in high dimensions, we hypothesize that theory analysis will demonstrate the first U-shaped and double descent will occur for pseudo-inverse LDA in the over-parameterized regime. With knowledge of the condition of double descent's occurrence, we can make a better decision on how to apply LDA to high dimensional model. For example, in a real-world gene expression analysis, we may not need to conduct dimension reduction at the cost of losing valuable information and keep large amounts of genes as features despite a small sample size.

Thesis Structure

The rest of the thesis will be structured as follow. In Chapter 2, we will outline the LDA error expression to set up our theoretical analysis, as well as presenting literature review of previous results. In Chapter 3, we will present the key results and definitions in RMT to help develop the theoretical proof. In Chapter 4, we will present the main theoretical result and proof. In Chapter 5 we will show simulation results of LDA's error rate against γ under different settings. In Chapter 6, we will conduct a real data analysis on cancer classification data. Key contributions and future work is discussed in chapter 7.

2 Error of Linear Discriminant Analysis

To evaluate the performance of a model, one must understand its error rate. Regression models have the advantage in this respect because their error has closed form expression composing of bias and variance. For example, if we assume a linear model $Y = f(X) + \varepsilon$, $\mathbb{E}[\varepsilon] = 0$, the expected prediction error of an unseen test point $X = X_0$ can be written as:

$$\text{Err}(X_0) = \mathbb{B}[(Y - \widehat{f}(x_0))^2] = \text{Bias}[\widehat{f}(x_0)]^2 + \text{Var}[\widehat{f}(x_0)] + \sigma_\varepsilon^2 \tag{2.1}$$

where σ^2 refers to irreducible error. Bias of the model is the amount which the average of the estimate differs from the truth: $\mathbb{E}\widehat{f}(x_0) - f(x_0)$. Variance is the expected squared deviation of the prediction around its mean: $\mathbb{E}[\widehat{f}(x_0) - \mathbb{E}\widehat{f}(x_0)]^2$ (Hastie et al., 2009). Since the predicted values of regression model is continuous, the analytical decomposition of the mean squared error (MSE) into these components is possible.

However, for classification methods, closed-form expression of bias-variance decomposition is often unavailable as the response is discrete and does not have a meaningful scale like continuous outputs. Error analysis then becomes more difficult. For example, Montanari et al. (2019) resorted to finding the upper and lower error bound of the maximum margin classifier. Different from other classification methods, linear discriminant analysis has a closed-form error expression McLachlan (1992). This facilitates the derivation of asymptotic results.

2.1 LDA Error Expression

Recall the setting of two class binary LDA classification with equal sample size and group prior ($n_1 = n_2, \pi_1 = \pi_2$). Let X be a p dimensional normal random vector with the distribution $X_{ki} \sim N(\mu_k, \Sigma)$, $i = 1, 2, \ldots n_k$, $k = 1, 2$. We can express the decision rule as:

$$\xi(X) = \mathbb{I}\Big\{(X - \frac{\mu_1 + \mu_2}{2})^\top \Sigma^{-1}(\mu_1 - \mu_2) > 0\Big\} \tag{2.2}$$

where $\mathbb{I}$ is an indicator function which assigns X to the group $k = 1$ if the expression $\xi(X) = 1$. Replacing the population parameters with estimates, the rule becomes:

$$\xi(X; \bar{X}_1, \bar{X}_2, S) = (X - \frac{\bar{X}_1 + \bar{X}_2}{2})^\top S^{-1}(\bar{X}_1 - \bar{X}_2) \tag{2.3}$$

The probability, conditional on the data, that a randomly chosen data point X from group $G_k, k = 1, 2$ is mis-allocated by the decision rule is the mis-classification error rate R_{LDA}:

$$ec_{12} = \mathbb{P}(\xi(X; \bar{X}_1, \bar{X}_2, S) < 0 | X \in G_1) \tag{2.4}$$

$$ec_{21} = \mathbb{P}(\xi(X; \bar{X}_1, \bar{X}_2, S) > 0 | X \in G_2) \tag{2.5}$$

$$R_{LDA} = \mathbb{P}(X \in G_1)ec_{12} + \mathbb{P}(X \in G_2)ec_{21} = \frac{1}{2}ec_{12} + \frac{1}{2}ec_{21} \tag{2.6}$$

The decision rule can be written as a linear form of X:

$$\xi(X; \bar{X}_1, \bar{X}_2, S) = \hat{\beta}_0 + \hat{\beta}^\top X \tag{2.7}$$

$$\hat{\beta}_0 = -\frac{1}{2}(\bar{X}_1 + \bar{X}_2)^\top S^{-1}(\bar{X}_1 - \bar{X}_2) \tag{2.8}$$

$$\hat{\beta} = S^{-1}(\bar{X}_1 - \bar{X}_2) \tag{2.9}$$

Since the data X is normally distributed, the distribution of $\xi(X; \bar{X}_1, \bar{X}_2, S)$ is also a transformed normal within group G_k, with group specific mean and common variance:

$$\mu_{\xi_k} = \hat{\beta}_0 + \hat{\beta}^\top \mu_k, \quad k = 1, 2$$
$$\Sigma_\xi = \hat{\beta}^\top \Sigma \hat{\beta}$$
$$\xi_k \sim N(\mu_\xi, \Sigma_\xi) \quad \xi_k \sim N(\hat{\beta}_0 + \hat{\beta}^\top \mu_k, \hat{\beta}^\top \Sigma \hat{\beta})$$

If we transform the distribution of allocation rule by subtracting the mean and dividing by the variance, the probability of misclassification for each group can be expressed as:

$$ec_i = \Phi\Big\{(-1)^k \frac{(\hat{\beta}_0 + \hat{\beta}^\top \mu_k)}{(\hat{\beta}^\top \Sigma \hat{\beta})^{\frac{1}{2}}}\Big\} \quad (k = 1, 2) \tag{2.10}$$

where $\Phi(\cdot)$ is the cumulative standard normal distribution.

$$\phi(z) = \frac{1}{\sqrt{2\pi}} exp(-\frac{z^2}{2})$$
$$\Phi(z) = \text{Prob}[Z \leq z] = \int_{-\infty}^{z} \phi(u)du = \frac{1}{2}\left[1 + erf(\frac{z}{\sqrt{2}})\right]$$

Z is the standard normal variable. We sum the probabilities to obtain the combined error rate for the two groups.

Lemma 2.1. *Conditional error rate of Linear Discriminant Analysis*

$$R_{LDA} = \sum_{k=1}^{2} \pi_k \Phi\Big\{\frac{(-1)^k\{\mu_k - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)\}^\top S^{-1}(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^\top S^{-1} \Sigma S^{-1}(\bar{X}_1 - \bar{X}_2)}}\Big\} \quad (k = 1, 2) \tag{2.11}$$

$\Phi(\cdot)$ *denotes cumulative standard normal distribution*

If we substitute sample estimates for the population parameters μ_k and Σ and use normal inverse, the mis-classification rate for each group becomes:

$$
\begin{aligned}
&\Phi\Big\{\frac{\{(\bar{X}_1-\frac{1}{2}(\bar{X}_1+\bar{X}_2)\}^\top S^{-1}(\bar{X}_1-\bar{X}_2)}{\sqrt{(X_1-\bar{X}_2)^\top S^{-1}SS^{-1}(\bar{X}_1-\bar{X}_2)}}\Big\}\\
=&\Phi\Big\{\frac{\{(\bar{X}_1-\frac{1}{2}(\bar{X}_1+\bar{X}_2)\}^\top S^{-1}(\bar{X}_1-\bar{X}_2)}{\sqrt{(\bar{X}_1-\bar{X}_2)^\top S^{-1}(\bar{X}_1-\bar{X}_2)}}\Big\}\\
=&\Phi\Big\{\frac{\frac{1}{2}(\bar{X}_1-\bar{X}_2)^\top S^{-1}(\bar{X}_1-\bar{X}_2)}{\sqrt{(\bar{X}_1-\bar{X}_2)^\top S^{-1}(\bar{X}_1-\bar{X}_2)}}\Big\}
\end{aligned}
$$

This is the Bayes error rate:

$$\Phi\Big(-\frac{1}{2}\Delta\Big) \tag{2.12}$$

where Δ is the estimated Mahalanobis distance between groups.

$$\Delta=\sqrt{(\bar{X}_1-\bar{X}_2)^\top S^{-1}(\bar{X}_1-\bar{X}_2)} \tag{2.13}$$

When $\Delta\to 0$, the two classes are indistinguishable and LDA is no better than random guessing (. When $\Delta\to\infty$, the two classes are so separate that misclassfication rate would be 0. It is also well-known that the sample LDA error rate is consistent with the Bayes rule as $n\to\infty$ when p is fixed. Therefore it achieves the minimum mis-classification rate among all classifiers under the homoscedastic Gaussian assumption Hamsici and Martinez (2008). Shao et al. (2011) also showed that as p diverges at a rate slower than n, $p\sqrt{\log p}/\sqrt{n}\to 0$, LDA error is also consistent with the optimal Bayes rule.

2.2 Review of Works on LDA Error Expression

Historically, there has been a line of work on deriving error estimators for LDA as well as finding out its behaviour under different settings.

Šarūnas Raudys and Young (2004) gave a summary of the theoretical work regarding the convergence of discriminant analysis from the Soviet Union. Okamoto (1963) was the first to

derive an asymptotic expansion of the expected error rate for sample linear discriminant function, under the setting of fixed p and $n \to \infty$. Hills (1966) considered the estimation of the error rate's expectation, including maximum likelihood, unbiased estimate, as well as its behaviour under different data distribution. Raudys (1967) was the first to use the double asymptotic approach to approximate the error rate of LDA with a known isotropic covariance matrix $\Sigma = I$. Deev (1970) formalized the double asymptotic approach and obtained asymptotic expansions for Gaussian and non-Gaussian models. Wyman et al. (1990) gave a comparison of asymptotic error rate expansions and found that they often yield good approximations even when training-sample sizes are small. Certain error estimators perform best when $n/p \leq 3$ whilst others favour small values of Mahalanobis distance between two populations. Asymptotic approximations of the expected probabilities of error rate when both dimension and sample size is large has also been studied by Fujikoshi (2000).

Comprehensive summary of error rate of discriminant analysis along with its estimators has been detailed by McLachlan (1992). Comparing the different error estimators, Zollanvari et al. (2011) studied the performance of error estimators for LDA by deriving the double asymptotic analytic expression for the first moments, second moments of the re-substitution error and leave-one-out estimator, and found that the re-substitution error estimator has an asymptotic bias of 0.

Other related work also involves with the variations of LDA discriminant rule (Saranadasa, 1993), the expected error of quadratic discriminant analysis in high dimensions (Cheng, 2004) and the asymptotic error rate of other classification criterion in high dimensions (Li and Yao, 2016).

A closely related work to ours is that of Bickel and Levina (2004). It is deduced that the worst-case performance of the misclassification rate of LDA using pseudo-inverse as $p/n \to \infty$ approaches $\frac{1}{2}$, which is equal to random guessing. The authors also investigated the error diagonal LDA where the covariance matrix is estimated by assuming independence of components by replacing off-diagonal elements with zeros. If the covariance matrix has eigenvalues going to 0 or ∞ as $p \to 0$, then the diagonal LDA also converges to random guessing. Otherwise, it

is superior than the pseudo-inverse approach when p increases faster than n.

Raudys and Duin (1998) derived the asymptotic formula for the expected error of the classifier with a different pseudo-inversion of the covariance matrix. The principal component is applied to the non-zero eigenvalues of the sample covariance matrix and the number of samples is increased from 1 to the p. The expected generalization error first decreases until obtaining minimum at $p = \frac{n}{4}$ then begins to increase again till n approaches p. This corresponds to the first descent of the double descent phenomenon in the under-parameterized region. Raudys and Duin (1998) also considered the over-parameterized case and found that expected error increasingly decrease with p. However, this result is not surprising due to the identity covariance matrix assumption.

Bian and Tao (2014) focused on the $p/n \to \gamma \in (0, 1)$ region and derived the bounds of both the discrimination power and generalization error. The result shows that both bounds are substantially determined by the γ: as γ increases, the upper bound of generalization error also increases, but n only needs to scale linearly relative to p to achieve acceptable performance.

Other derivations include Dobriban and Wager (2018) 's work on the closed-form expression of the limiting predictive risk of Regularized Discriminant Analysis (RLDA) in the high dimensional $\gamma > 0$ setting, with the assumption that each predictor has a small, independent random effect on the outcome. It was found that the asymptotic of RLDA can be expressed in terms of the angle between the Bayes decision boundary hyperplane and RLDA discriminating plane, which tends to be an asymptotically deterministic value in terms of the covariance matrix. Wang and Jiang (2018) also derived the error of regularized discriminant analysis but without the random effect assumption.

Recent work by Cheng et al. (2022) studied the effects of imbalanced data on LDA, called label shift, where training data for each group is no longer equal due to sampling bias. It is shown that the asymptotic error of LDA trained on imbalanced data can outperform the reduced sample balanced data model at certain values of γ. Down-sampling, or reducing training data so that class proportions are equal, may hurt model performance; but when the data imbalance is severe, it can be beneficial.

Wang and Jiang (2018) investigated the effect on LDA error rates in the setting of diverging p and $\gamma < 1$, considering cases when either the covariance matrix is known or the group mean is known. Results are derived under the assumption of Gaussian distributed data and using properties of Wishart distribution. We present the results here as a preliminary guide to understanding the effects of LDA error under either unknown covariance matrix Σ or group mean μ_k.

Lemma 2.2 (Theorem 2.1 of Wang and Jiang (2018))**.** *Let $p/n_k \to \gamma_k \in (0, \infty)$, $k = 1, 2$ also assuming known Σ, the misclassification rate of LDA with Bayes classifier is:*

$$R_{LDA1} \xrightarrow{p} \frac{1}{2}\Phi\left(-\frac{\Delta^2 + \gamma_2 - \gamma_1}{2\sqrt{\Delta^2 + \gamma_1 + \gamma_2}}\right) + \frac{1}{2}\Phi\left(-\frac{\Delta^2 + \gamma_1 - \gamma_2}{2\sqrt{\Delta^2 + \gamma_1 + \gamma_2}}\right)$$

where $\Delta = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)}$.

When Σ is known and μ_k is unknown, as γ_k increases, the classification rate exceeds Bayesian error and tends towards $1/2$, which is similar to random guessing.

Lemma 2.3 (Theorem 2.2 of Wang and Jiang (2018))**.** *Assuming $p/(n_1 + n_2) \to \gamma \in (0, 1)$, and the population means are μ_1, μ_2 are known. The sample covariance matrix is defined as $S_n = \frac{1}{n-2}\sum_{k=1}^{2}\sum_{j=1}^{n_k}(X_{kj} - \bar{X}_k)(X_{kj} - \bar{X}_k)^\top$. Then the LDA error approaches:*

$$R_{LDA2} \xrightarrow{p} \Phi(-\frac{\Delta}{2}\sqrt{1-\gamma})$$

where $\Delta = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)}$.

Again, when Σ is unknown, the error rate has the additional term $\sqrt{1-\gamma}$ compared to the Bayes error. However, when γ is small, the increase in Δ increases discrimination power and decreases error. However, when γ reaches beyond a point, the estimation price we pay for not knowing both the Σ and μ_k increases with γ, which causes the peaking phenomenon and the first U-shaped curve.

3 Foundation of Random Matrix Results

In this chapter, we present the mathematical preliminaries from Large Dimensional Random Matrix Theory (LDRMT) that are needed to develop the theoretical result. RMT is a branch of mathematical statistics that is concerned with the statistical properties of matrices with entries as random variables. It has origins in mid 20^{th} century in the field of nuclear physics, when Eugene Wigner introduced modelling energy levels of nuclei with large real symmetric random matrices Wigner (1951). Due to the universality property of random matrices, developments in RMT found applications in diverse areas including high-dimensional data analysis, wireless communication, finance and models of complex systems. One of the central focuses of RMT is to understand the eigenvalue distribution of random matrices. Key theorems include the semi-circle law (Wigner, 1955), M-P law (Marčenko and Pastur, 1967), Tracy-Widom distribution (Tracy and Widom, 1994). We will mainly present results pertaining to sample covariance matrices

3.1 Empirical Spectral Distribution (ESD)

A fundamental concept in RMT is the distribution of eigenvalues of random matrices. Given a $p \times p$ matrix $\mathbf{A}$, a scalar value λ is called the eigenvalue of A if there exists a nonzero vector $\mathbf{v}$, eigenvector such that

$$\mathbf{Av} = \lambda\mathbf{v} \tag{3.1}$$

Definition 3.1 (Empirical Spectral Distribution)**.** *For a $(p \times p)$ matrix A with real eigenvalues $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$, We define F^A as the empirical Spectral distribution of the eigenvalues,*

i.e.,

$$F^A(x) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}(\lambda_i \leq x), \quad x \in \mathbb{R} \tag{3.2}$$

where $\mathbb{I}(\cdot)$ *is the indicator function.*

Another useful definition is the spectral decomposition, which can be written as follows:

Definition 3.2 (Spectral Theorem)**.** *For any symmetric matrix* $A \in \mathbb{R}^{p\times p}$ *with eigenvalues* $\lambda_1, \lambda_2, \ldots, \lambda_p \in \mathbb{R}$ *and corresponding eigenvectors* $v_1, v_2, \ldots, v_p$*, then* A *can be written as:*

$$A^k = \sum_{i=1}^{n} \lambda_i^k v_i v_i^\top$$

3.2 Limiting Spectral Distribution (LSD)

A central issue in RMT is to find the convergence of ESD of a sequence of random matrices $\{S_n\}$, which is the limiting spectral distribution (LSD). In the classical under-parametrized setting, as $n \to \infty$ and p is fixed, the eigenvalues of the sample covariance matrix S converge to their population counterpart almost surely, due to the law of large numbers. For example, in the case where the population covariance matrix is the identity $\Sigma = \mathbb{I}$, the ESD of the sample covariance matrix S becomes more and more narrowly centered around one as n increases. However, when p increases proportionally, the distribution moves to the left quickly and cannot be approximated by a centered normal distribution (Bai et al., 2007).

This phenomenon motivated the study of matrices with the number of columns tending to infinity and the limiting properties of the sample covariance matrix ESD. It was then discovered that, even under the high dimensional asymptotic setting, ESD could converge to a deterministic distribution function.

Definition 3.3 (Limiting Spectral Distribution)**.** *For a sequence of* $p\times p$ *random matrices* $\{S_n\}$*, if the sequence of corresponding ESD* $\{F^{S_n}\}$ *vaguely converges to a distribution* F *, and* F *is a proper c.d.f., then* F *is the limiting spectral distribution (LSD) of* $\{S_n\}$*. That is, for any* φ

continuous function that is compactly supported :

$$\int_{-\infty}^{\infty} \varphi(x) F^{S_n}(x) dx \to \int_{-\infty}^{\infty} \varphi(x) F(x) dx$$

3.3 Marchenko-Pastur Law

The much celebrated Marchenko-Pastur law (Marčenko and Pastur, 1967) characterizes the limiting distribution of the eigenvalues of the sample covariance matrix when the population covariance matrix is a multiple of the identity.

Lemma 3.4 (Marčenko and Pastur (1967))**.** *Suppose that X is a rectangular matrix with entries as random variables $X = \{X_{ij}\} \in \mathbb{R}^{n \times p}, \mathbb{E}[X_{ij}] = 0, \mathbb{E}[X_{ij}^2] = \sigma^2$. Let the sample covariance matrix be $S_n = \frac{1}{n} X X^\top$ and define the spectral measure $F^{S_n} = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i}$. Also assume that $p/n \to \gamma \in (0, \infty)$ as $n, p \to \infty$. With probability one, F^{S_n} converges to a deterministic distribution with density*

$$p_\gamma(x) = \begin{cases} \frac{1}{2\pi\sigma^2 x\gamma} \sqrt{(b-x)(x-a)}, & \text{if } a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

where $a(\gamma) = \sigma^2(1 - \sqrt{\gamma})^2, b(\gamma) = \sigma^2(1 + \sqrt{\gamma})^2$. The density has point mass $(1 - 1/\gamma)$ at the origin if $\gamma > 1$.

This result implies that the largest eigenvalue of the population covariance matrix λ_p is overestimated by the largest value of the sample covariance matrix. The smallest eigenvalue of the population covariance matrix λ_1 is underestimated by the smallest eigenvalue of the sample covariance matrix. The eigenvalues no longer concentrate at their true value but span a range around it. Therefore, the estimation price we pay for the population covariance matrix increases when p becomes comparable to n.

3.4 Stieltjes Transform

Another important tool in RMT is the Stieltjes transform, which can be used to characterize the LSD.

Definition 3.5. *The Stieltjes transform $m_F(z)$ for a function of bounded variation $F(x)$ is defined as*

$$m_F(z) \equiv \int \frac{1}{x-z} dF(x), \quad z \in \mathbb{C}^+ \equiv \{z \in \mathbb{C} : \Im z > 0\}$$

The significance of Stieltjes transform is that the convergence of ESDs can be found through the convergence of their Stieltjes transforms. Hence, the transform acts similarly as characteristic functions that help to identify distributions of eigenvalues, which is evident from the below result:

Lemma 3.6 (Theorem B.2.1 of Bai and Silverstein (2010))**.** *Assuming that $\{G_n\}$ is a sequence of functions of bounded variations with $G_n(-\infty) = 0$ for all n. Then,*

$$\lim_{n\to\infty} m_{G_n}(z) = m(z) \quad \forall z \in \mathbb{C}^+ \tag{3.4}$$

if and only if there is a function of bounded variation G with $G(-\infty) = 0$ and Stieltjes transform $s(z)$ and such that $G_n \to G$ vaguely.

An important result applied the Stieltjes transform to characterize the LSD of a matrix:

Lemma 3.7 (Theorem 4.1 of Bai and Silverstein (2010) and Theorem 1.1 of Silverstein (1995))**.** *Suppose that the entries of $(n \times p)$ matrix $X_n = (X_{ij}^n)$, $(n \times p)$ are independent complex random variables satisfying $\mathbb{E}|X_{11}^1 - \mathbb{E}X_{11}^1|^2 = 1$. T_n is a sequence of Hermitian matrices independent of X_n and the ESD of T_n , F^{T_n} converges almost surely in distribution to a p.d.f H on $[0, \infty)$. let $B_n = \frac{1}{n} X_n^* T_n X_n$. If $p/n \to \gamma \in (0, \infty)$, as $n, p \to \infty$, then F^{B_n}, the ESD of B_n, converges almost surely to a nonrandom p.d.f F whose Stieltjes transform $m_F(z)$ is the solution to the*

equation:

$$m_F(z) = \int \frac{1}{t(1-\gamma-\gamma z m_F(z)) - z} dH(t) \tag{3.5}$$

which is unique in the set $\{m_F(z) \in \mathbb{C} : -(1-\gamma)/z + \gamma m_F(z) \in \mathbb{C}^+\}$.

3.5 High-Dimensional Hotelling T^2 Statistics

The Hotelling T^2 distribution is a generalization of the Student's t-distribution for hypothesis testing of the means of the population in the multivariate case (Hotelling, 1931). Assume that $X_1, ..., X_{n_x} \sim N_p(\mu, \Sigma)$, let sample means be $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and sample covariance matrices be $S = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top$. Then the Hotelling's T^2 statistics for testing the null hypothesis $H_0 : \mu = 0$ against $H_A : \mu \neq 0$ is defined as $t^2 = (\bar{X} - \mu)^\top S^{-1}(\bar{X} - \mu)$. It follows F distribution with parameters $p, n-p$: $\frac{p(n-1)}{n-p} F_{p,n-p}$.

However, as p becomes larger compared to sample size n, the classical test fails due to changes in sampling distribution (Paul and Aue, 2014). When $p > n$, the sample covariance matrix is also singular and not invertible. Pan and Zhou (2011) investigated the limiting distribution of the test statistics under the double asymptotic setting and derived the following result:

Lemma 3.8 (Theorem 2 of Pan and Zhou (2011))**.** *Let* $s_j = (X_{1j}, \ldots, X_{pj})^\top$*, where* $\{X_{ij}\}$ *is a double array of i.i.d real random variables with* $E[X_{ij}] = 0$*,* $E[X_{ij}^2] = 1$ *and* $E[X_{ij}^4] < \infty$*. Let the sample covariance matrix be* $S = \frac{1}{n}\sum_{j=1}^n (s_j - \bar{s})(s_j - \bar{s})^\top$*. Suppose* $p \leq n, \gamma_n = p/n \to \gamma > 0$ *as* $n \to \infty$*.* $g(x)$ *is a function with a continuous first derivative in a neighborhood of* γ *and* $f(x)$ *is analytic on an open region containing the interval* $[I_{(0,1)}(\gamma)(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$*. Then,*

$$\left(\sqrt{n}\Big[\frac{\bar{s}^\top f(S)\bar{s}}{||\bar{s}||^2} - \int f(x) dF_\gamma(x)\Big], \sqrt{n}(g(\bar{s}^\top \bar{s}) - g(\gamma))\right) \xrightarrow{D} (X, Y),$$

where F_γ *is the Marchenko-pastur law,* X *is a Gaussian random variable with* $E(X) = 0$ *and* $Y \sim N(0, 2c(g'(c))^2)$.

3.6 Results Relating to Eigenvectors

There has been an abundance of literature on the distribution of eigenvalues of random matrices. The sure convergence of the largest and smallest eigenvalue of the sample covariance matrix have both been investigated (Yin et al., 1988; Bai and Yin, 1993). In terms of the behaviour of eigenvectors, the work has been more limited. Bai et al. (2007) investigated the limiting behaviour of matrix S_n by defining a new form of ESD with weights determined by eigenvectors. The authors proved that its limiting distribution is the same as the ESD defined by equal weights. Let $U_n\Lambda_n U_n^*$ be the spectral decomposition of S_n where $U_n = (u_{ij})$ is a unitary matrix consisting of orthonormal eigenvectors. For an arbitrary nonrandom unit vector $\mathbf{x}_n$, $U_n^*\mathbf{x}_n = (y_1, y_2, \ldots, y_n)^\top = \mathbf{y}$. $\mathbf{y}$ has unit norm and if U_n is asymptotically Haar distributed then $\mathbf{y}$ should be asymptotically uniformly distributed over the unit sphere. The authors define a new empirical spectral distribution of S_n :

$$F_1^{S_n}(x) = \sum_{i=1}^{n} |y_i|^2 \mathbb{I}(\lambda_i \leq x)$$

The Stieltjes transform of $F_1^{S_n}(x)$ is given by

$$m_{F_1^{A_n}}(z) = x_n^*(S_n - zI)^{-1}x_n = \int \frac{1}{x-z} dF_1^{S_n}(x) \tag{3.6}$$

The following result finds the LSD of the new spectral distribution:

Lemma 3.9 (Theorem 1 of Bai et al. (2007))**.** *1. Let $X_n = (X_{ij})$ be an $n \times p$ matrix of complex random variable with $\mathbb{E}[X_{ij}] = 0$, $\mathbb{E}[X_{ij}]^2 = 1$ and $\mathbb{E}[X_{ij}]^4 < \infty$. Let $S_n = \frac{1}{N} T_n^{1/2} X_n X_n^* T_n^{1/2}$, where T_n is a non-negative definite matrix with its spectral norm bounded in p.*

2. *Let $\boldsymbol{x}_n \in \mathbb{C}_1^n = \{\boldsymbol{x} \in \mathbb{C}^n, ||\boldsymbol{x}|| = 1\}$*

3. *$\frac{p}{n} \to \gamma \in (0, \infty)$ as $n, p \to \infty$.*

4. *If* $H_n = F^{T_n} \xrightarrow{\mathbb{D}} H$ *is a proper distribution function and* $\boldsymbol{x}_n^*(T_n - zI)^{-1}\boldsymbol{x}_n \to m_{F^H}(z)$, *where* $m_{F^H}(z)$ *denotes the Stieltjes transform of* $H(t)$. *Then*

$$F_1^{S_n}(x) \xrightarrow{a.s.} F^{\gamma,H}(x) \tag{3.7}$$

where $F^{\gamma,H}(x)$ *is the limiting spectral distribution of* S_n.

Note that the condition $\mathbf{x}_n^*(T_n - zI)^{-1}\mathbf{x}_n \to m_{F^H}(z)$ holds if T_n is a positive multiple of the identity matrix. More generally, if $\lambda_{max}(T_n) - \lambda_{min}(T_n) \to 0$, then the condition holds uniformly for all $\mathbf{x}_n \in \mathbb{C}_1^n$. The condition is also satisfied if the eigenvector matrix of a sample covariance matrix transforms $\mathbf{x}_n$ to a unit vector whose entries' absolute values are close to $1/\sqrt{N}$. For example, if $\mathbf{x}_n = (\mathbf{u}_1 + \cdots + \mathbf{u}_n)/\sqrt{n}$, where $(\mathbf{u}_1, \ldots, \mathbf{u}_n)$ are orthonoromal eigenvectors of the spectral decomposition of T_n.

One of the applications of RMT is in mean-variance portfolio optimization (Markowitz, 1952). The problem involves investors choosing a set of assets such that the portfolio guarantees a level of return but the variance, or the risk associated, is minimized. Using RMT, Bai et al. (2009) demonstrates that the estimated optimal return in high dimensional setting is always $\sqrt{\frac{1}{1-\gamma}}$ times larger than the theory prediction. Similarly, Karoui (2010) found that the price of estimating the covariance matrix translates to underestimating the variance by roughly $1-\gamma$. Both these works deal with the sample estimate of the covariance matrix and its quadratic form's convergence. Since covariance matrix estimation is also involved in LDA error, we will draw from these existing results.

4 Main Results

4.1 Under-parameterized Regime

For this chapter, we will present the result for the asymptotic convergence of the LDA error rate under double asymptotic setting $p \to \infty$. $n \to \infty$, $p/n \to \gamma \in (0, 1)$. We will develop the result based on the closed form expression of LDA error rate:

$$R_{\text{LDA}} = \frac{1}{2}\sum_{k=1}^{2} \Phi\{\frac{(-1)^k\{\mu_k - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)\}^\top S^{-1}(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^\top S^{-1}\Sigma S^{-1}(\bar{X}_1 - \bar{X}_2)}}\} \quad (k = 1, 2) \tag{4.1}$$

Assumption 1. To develop the proof we will assume the following data structure for X. $\{X_{1i}, i = 1, \ldots, n_1\}$ and $\{X_{2i}, i = 1, \ldots, n_2\}$ are independent and identically distributed:

$$X_{ki} = \mu_k + Z_{ki} \quad Z_{ki} = \Sigma^{\frac{1}{2}} Y_{ki} \quad k = 1, 2; \quad i = 1, \ldots n_k \tag{4.2}$$

$$\mathbb{E}[Y_i] = 0, \quad \mathbb{E}[Y_i^2] = 1, \quad \mathbb{E}[Y_i^4] < \infty \tag{4.3}$$

We will assume equal group samples and $n_1 = n_2$. We estimate μ_1, μ_2 and Σ by their sample analogs. (Note that for ease if calculation we will not use bias corrected version for pooled sample covariance matrix)

$$\bar{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki}, \quad k = 1, 2 \tag{4.4}$$

$$S = \frac{1}{2n_1}\sum_{i=1}^{n_1}(X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1)^\top + \frac{1}{2n_2}\sum_{i=1}^{n_2}(X_{2i} - \bar{X}_2)(X_{2i} - \bar{X}_2)^\top \tag{4.5}$$

We present the result of LDA's error convergence in the under-parameterized regime.

Theorem 1. *Following the assumption 1 and the following conditions:*

1. $\frac{p}{n} \to \gamma \in (0,1)$ *as* $n \to \infty$
2. $\frac{\mu_1^\top \Sigma^{-1} \mu_1}{n} \to a_1$, $\frac{\mu_2^\top \Sigma^{-1} \mu_2}{n} \to a_2$, $\frac{\mu_1^\top \Sigma^{-1} \mu_2}{n} \to a_3$, $\frac{(\mu_1+\mu_2)^\top \Sigma^{-1}(\mu_1+\mu_2)}{n} \to b_1$, $\frac{(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2)}{n} \to b_2$, $\quad a_1, a_2, a_3, b_1, b_2 \neq 0$

The error rate of the linear discriminant analysis in equal sample size setting converges in probability to the following.

$$R_{LDA} \xrightarrow{a.s.} \Phi\Big\{-2\Delta_* \sqrt{\gamma(1-\gamma)}\Big\}$$

Where $\Delta_*^2 = \lim_{n\to\infty, p\to\infty} \frac{(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2)}{n}$

4.2 Proof Outline

The proof method for Theorem 1 involves expanding the numerator and denominator term inside $\Phi(\cdot)$ and finding their convergence. We mainly leverage existing random matrix result from Bai et al. (2009), Karoui (2010) to find the convergence of the quadratic terms involving S^{-1}, $\Sigma S^{-1} \Sigma$ with either deterministic or random vectors on each-side.

A note on the sample covariance matrix

Many of the techniques used to prove the convergence of the error rate use the single group definition of the sample covariance matrix. However, the sample covariance matrix used in LDA is the average of two groups. To see how we can apply the previous results, we decompose the $X_{ki} - \bar{X}_k$ into $\bar{Z}$ and pool the two groups together so that we treat all observations from two groups as as one group $i = 1, \ldots, n$. We then have leftover terms $\frac{1}{4}(\bar{Z}_1 \bar{Z}_1^\top + \bar{Z}_2 \bar{Z}_2^\top + 2\bar{Z}_1 \bar{Z}_2^\top)$. This term is low rank and would not affect the convergence in our asymptotic setting. Thus, we treat the sample covariance as $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and apply previous results.

4.2.1 Numerator Convergence

Numerator Term 1

We first expand the first term in the numerator:

$$\begin{aligned}(\bar{X}_1 + \bar{X}_2)^T S^{-1}(\bar{X}_1 - \bar{X}_2) &= (\mu_1 + \bar{Z}_1 + \mu_2 + \bar{Z}_2)S^{-1}(\mu_1 + \bar{Z}_1 - \mu_2 - \bar{Z}_2) \\ &= \mu_1^\top S^{-1}\mu_1 + \mu_1^\top S^{-1}\bar{Z}_1 + \mu_1^\top S^{-1}\mu_2 - \mu_1^\top S^{-1}\bar{Z}_2 \\ &\quad + \bar{Z}_1^\top S^{-1}\mu_1 + \bar{Z}_1^\top S^{-1}\bar{Z}_1 - \bar{Z}_1^\top S^{-1}\mu_2 - \bar{Z}_1^\top S^{-1}\bar{Z}_2 \\ &\quad + \mu_2^\top S^{-1}\mu_1 + \mu_2^\top S^{-1}\bar{Z}_1 - \mu_2^\top S^{-1}\mu_2 - \mu_2^\top S^{-1}\bar{Z}_2 \\ &\quad + \bar{Z}_2^\top S^{-1}\mu_1 + \bar{Z}_2^\top S^{-1}\bar{Z}_1 - \bar{Z}_2^\top S^{-1}\mu_2 - \bar{Z}_2^\top S^{-1}\bar{Z}_2\end{aligned}$$

Since the covariance matrix S is symmetric and the inverse S^{-1} is also symmetric. This means that the certain terms cancel out and the expression can then be simplified to:

$$\begin{aligned}(\bar{X}_1 + \bar{X}_2)^\top S^{-1}(\bar{X}_1 - \bar{X}_2) &= \mu_1^\top S^{-1}\mu_1 - \mu_2^\top S^{-1}\mu_2 \\ &\quad + 2\mu_1^\top S^{-1}\bar{Z}_1 - 2\mu_2^\top S^{-1}\bar{Z}_2 \\ &\quad + \bar{Z}_1^\top S^{-1}\bar{Z}_1 - \bar{Z}_2^\top S^{-1}\bar{Z}_2 \qquad (4.6)\end{aligned}$$

Numerator Term 2

$$\mu_1^\top S^{-1}(\bar{X}_1 - \bar{X}_2) = \mu_1^\top S^{-1}\mu_1 + \mu_1^\top S^{-1}\bar{Z}_1 - \mu_1^\top S^{-1}\mu_2 - \mu_1^\top S^{-1}\bar{Z}_2 \qquad (4.7)$$

$$\mu_2^\top S^{-1}(\bar{X}_1 - \bar{X}_2) = \mu_2^\top S^{-1}\mu_1 + \mu_2^\top S^{-1}\bar{Z}_1 - \mu_2^\top S^{-1}\mu_2 - \mu_2^\top S^{-1}\bar{Z}_2 \qquad (4.8)$$

It is then necessary to find the convergence of the following terms for the numerator:

1.1. $\mu_1^\top S^{-1}\mu_1, \mu_2^\top S^{-1}\mu_2$

1.2. $\mu_1^\top S^{-1}\mu_2, \mu_2^\top S^{-1}\mu_1$

1.3. $\bar{Z}_1 S^{-1}\bar{Z}_1, \bar{Z}_2^\top S^{-1}\bar{Z}_2$

1.4. $\mu_1^\top S^{-1}\bar{Z}_1, \mu_2^\top S^{-1}\bar{Z}_2, \mu_2^\top S^{-1}\bar{Z}_1, \mu_1^\top S^{-1}\bar{Z}_2$

Convergence of term 1.1

Lemma 4.1. *Suppose we have $k = 2$ groups of observations X_{ki} with n_k samples in each group ($n_1 = n_2, n_1 + n_2 = n$), $X_{ki} = \mu_k + Z_{ki}$, Where $Z_{ki} = \Sigma^{\frac{1}{2}} Y_k$, and $S = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^\top$. If $p/n \to \gamma \in (0, 1)$*

$$\frac{\mu_k^\top S^{-1}\mu_k}{n} \xrightarrow{a.s.} \frac{a_k}{1-\gamma} \quad \textit{for} \quad k = 1, 2$$

where $a_k = \lim_{n\to\infty, p\to\infty} \frac{\mu_k^\top \Sigma^{-1}\mu_k}{n}$, $a_k \neq 0$, for $k = 1, 2$

Proof. The proof follows from Lemma A.2 of Bai et al. (2009) with modification. Suppose $\frac{\mu_k^\top \Sigma^{-1}\mu_k}{n} \to a_k$, let $\tilde{\mu}_k = \Sigma^{-1/2}\mu_k$ and $\alpha_k = \frac{\tilde{\mu}_k}{||\tilde{\mu}_k||}$ is a nonrandom unit vector

$$\mu_k^\top S^{-1}\mu_k = ||\tilde{\mu}_k||^2 \alpha_k^\top \tilde{S}^{-1}\alpha_k$$
$$\text{where} \quad \tilde{S} = \Sigma^{-1/2} S \Sigma^{-1/2}$$

Using Corollary 2 of Bai et al. (2007) , we apply the convergence result using Stieljes transform of the Marchenko-Pastur law of $\tilde{S}_*$'s LSD F_γ:

$$\lim_{n\to\infty} \alpha_k^\top \tilde{S}^{-1}\alpha_k \xrightarrow{a.s.} \int x^{-1} dF_\gamma(x) = \frac{1}{1-\gamma}$$
$$\frac{1}{n}\mu_k^\top \Sigma^{-1}\mu_k = \frac{1}{n}||\tilde{\mu}_k||^2 \to a_k$$

Thus,

$$\frac{1}{n}||\tilde{\mu}_k||^2 \alpha_k^\top \tilde{S}^{-1}\alpha_k = \frac{\mu_k^\top S^{-1}\mu_k}{n} \xrightarrow{a.s.} \frac{a_k}{1-\gamma}$$

In our case, the single group S_* should be transformed to the pooled definition of the sample covariance matrix. We multiply the limit by 2. □

Convergence of Term 1.2

Lemma 4.2. *Using the assumptions in lemma 4.1, If* $p/n \to \gamma \in (0,1)$ *then we have*

$$\frac{\mu_1^\top S^{-1}\mu_2}{n} \xrightarrow{a.s.} \frac{a_3}{1-\gamma}$$

where $a_3 = \lim_{n\to\infty, p\to\infty} \frac{\mu_1^\top \Sigma^{-1}\mu_2}{n}$, $a_3 \neq 0$ *By symmetry, we also have that*

$$\frac{\mu_2^\top S^{-1}\mu_1}{n} \xrightarrow{a.s.} \frac{a_3}{1-\gamma}$$

Proof. Assuming that

$$\frac{\mu_1^\top \Sigma^{-1}\mu_2}{(\mu_1+\mu_2)^\top \Sigma^{-1}(\mu_1+\mu_2)} \text{ and } \frac{\mu_1^\top \Sigma^{-1}\mu_2}{(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2)}$$

Are bounded away from zero and apply the following decomposition:

$$\begin{aligned} 4\frac{\mu_1^\top S^{-1}\mu_2}{\mu_1^\top \Sigma^{-1}\mu_2} &= \frac{(\mu_1+\mu_2)^\top S^{-1}(\mu_1+\mu_2)}{(\mu_1+\mu_2)^\top \Sigma^{-1}(\mu_1+\mu_2)} \frac{(\mu_1+\mu_2)^\top \Sigma^{-1}(\mu_1+\mu_2)}{\mu_1^\top \Sigma^{-1}\mu_2} \\ &\quad - \frac{(\mu_1-\mu_2)^\top S^{-1}(\mu_1-\mu_2)}{(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2)} \frac{(\mu_1-\mu_2)^\top \Sigma^{-1}(\mu_1-\mu_2)}{\mu_1^\top \Sigma^{-1}\mu_2} \\ \frac{\mu_1^\top S^{-1}\mu_2}{\mu_1^\top \Sigma^{-1}\mu_2} &\to \frac{n}{1-\gamma} \end{aligned} \tag{4.9}$$

Thus, we have the desired result.

□

Convergence of Term 1.3

Lemma 4.3. *Using the assumptions of previous lemmas,*

$$\frac{\bar{Z}_k^\top S^{-1}\bar{Z}_k}{n} \xrightarrow{a.s.} 0 \quad k = 1,2$$

Proof. We follow Lemma 3.1 of Bai et al. (2009) and modify $\bar{Z}$ to one group case to obtain the

same convergence result. □

Convergence of Term 1.4

Lemma 4.4. *Using the assumptions of the previous lemma*

$$\frac{\mu_1^\top S^{-1}\bar{Z}_1}{n} \xrightarrow{a.s.} 0 \quad k = 1, 2 \tag{4.10}$$

Likewise $\frac{\mu_2^\top S^{-1}\bar{Z}_2}{n} \to 0$, $\frac{\mu_2^\top S^{-1}\bar{Z}_2}{n} \to 0$, $\frac{\mu_1^\top S^{-1}\bar{Z}_2}{n} \to 0$, $\frac{\mu_2^\top S^{-1}\bar{Z}_1}{n} \to 0$

The proof follows from applying Cauchy inequality

$$\frac{\mu_1^\top S^{-1}\bar{Z}_1}{n} \leq \sqrt{\frac{\mu_1^\top S^{-1}\mu_1}{n}}\sqrt{\frac{\bar{Z}_1^\top S^{-1}\bar{Z}_1}{n}}$$

$$\text{since} \quad \sqrt{\frac{\bar{Z}_1^\top S^{-1}\bar{Z}_1}{n}} \xrightarrow{a.s.} 0$$

$$\frac{\mu_1^\top S^{-1}\bar{Z}_1}{n} \xrightarrow{a.s.} 0$$

Similar logic can be applied to other terms in 1.4.

4.2.2 Denominator Convergence

Similarly, for the denominator, we will focus on a term involving $S^{-1}\Sigma S^{-1}$, it is easy to see that $S^{-1}\Sigma S^{-1}$ is symmetric because $(S^{-1}\Sigma S^{-1})^\top = (S^{-1})^\top \Sigma^\top (S^{-1})^\top = S^{-1}\Sigma S^{-1}$, so some

terms in the expansion combine:

$$\begin{aligned}
&\sqrt{(\bar{X}_1 - \bar{X}_2)^\top S^{-1}\Sigma S^{-1}(\bar{X}_1 - \bar{X}_2)} \\
&= \mu_1^\top S^{-1}\Sigma S^{-1}\mu_1 + \mu_2^\top S^{-1}\Sigma S^{-1}\mu_2 \\
&+ \bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1 + \bar{Z}_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_2 \\
&- 2\mu_1^\top S^{-1}\Sigma S^{-1}\mu_2 \\
&+ 2\mu_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1 + 2\mu_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_2 \\
&- 2\bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_2 \\
&- 2\mu_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_2 - 2\mu_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_1
\end{aligned}$$

Therefore, the terms we need to find convergence are:

2.1 $\mu_1^\top S^{-1}\Sigma S^{-1}\mu_1, \mu_2^\top S^{-1}\Sigma S^{-1}\mu_2$

2.2 $\mu_1^\top S^{-1}\Sigma S^{-1}\mu_2$

2.3 $\bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_2, \bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1, \bar{Z}_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_2$

2.4 $\mu_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1, \mu_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_2, \mu_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_2, \mu_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_1$

Convergence of Term 2.1

Lemma 4.5. *Using the assumptions of the previous lemma*

$$\frac{\mu_k^\top S^{-1}\Sigma S^{-1}\mu_k}{n||\bar{Y}_1 + \bar{Y}_2||^2} \xrightarrow{a.s.} \frac{a_k}{(1-\gamma)^3} \quad \textit{for} \quad k = 1, 2$$

$a_k = \lim_{n\to\infty, p\to\infty} \frac{\mu_k^\top \Sigma^{-1}\mu_k}{n}$ *for* $k = 1, 2$

Proof. From Theorem 4.1 of Karoui (2010), we can rewrite the expression $X - \bar{X}$

$$X - \bar{X} = \big(1 - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n}\big)X = HX \tag{4.11}$$

$\mathbf{1}_n$ is a n-dimensional vector with entries 1. Then rewriting the sample covariance matrix with the H matrix:

$$S = \frac{1}{n}(HX)^\top(HX) = \frac{1}{n}X^\top HX \tag{4.12}$$

Note the property

$$H^\top H = I - 2\frac{\mathbf{1}_n\mathbf{1}_n^\top}{n} + \frac{\mathbf{1}_n\mathbf{1}_n^\top\mathbf{1}_n\mathbf{1}_n^\top}{n^2} = I - \frac{\mathbf{1}_n\mathbf{1}_n^\top}{n} = H \tag{4.13}$$

We can then change the expression for S and S_*^{-1}

$$S = \frac{1}{n}\Sigma^{1/2}Y^\top HY\Sigma^{1/2} \tag{4.14}$$

$$S^{-1} = \Sigma^{-1/2}\big(\frac{1}{n}Y^\top HY\big)^{-1}\Sigma^{-1/2} \tag{4.15}$$

The term $S^{-1}\Sigma S^{-1}$ can then be written as

$$S^{-1}\Sigma S^{-1} = \Sigma^{-1/2}\Big(\frac{1}{n}Y^\top HY\Big)^{-2}\Sigma^{-1/2} \tag{4.16}$$

Since $\mu_1^\top\Sigma^{-1}\mu_1$ can be written as $||\Sigma^{-1/2}\mu_1||_2^2$, we then have the following transformation:

$$\begin{aligned}\frac{\mu_1^\top S^{-1}\Sigma S^{-1}\mu_1}{\mu_1^\top\Sigma^{-1}\mu_1} &= \frac{\mu_1^\top\Sigma^{-1/2}}{||\mu_1^\top\Sigma^{-1/2}||_2}\Big(\frac{1}{n}Y^\top HY\Big)^{-2}\frac{\Sigma^{-1/2}\mu_1}{||\Sigma^{-1/2}\mu_1||_2} \\ &= \nu_1^\top\Big(\frac{1}{n}Y^\top HY\Big)^{-2}\nu_1\end{aligned} \tag{4.17}$$

Where $\nu_1 = \frac{\Sigma^{-1/2}\mu_1}{||\Sigma^{-1/2}\mu_1||_2}$ is a vector of l_2 norm 1 and H is low rank. We then apply the spectral theorem and Bai et al. (2007) 's result on limiting distribution of ESD involving eigenvectors

(Lemma 3.9).

$$\nu_1^\top \Big(\frac{1}{n} Y^\top H Y\Big)^{-2} \nu_1 = \sum_{i=1}^{p} \frac{1}{\lambda_i^2} (\nu^\top u_i)^2$$
$$= \int \frac{1}{x^2} dF_1^{S_n}(x) \xrightarrow{a.s.} \int \frac{1}{x^2} dF_\gamma(x)$$

Recall that F_γ is the standard Marchenko-Pastur distribution. Solving the integral (see Appendix) by applying Stieljes transform of the M-P law, we obtain the result

$$\int \frac{1}{x^2} dF_1^{S_n}(x) = \int_{(1-\sqrt{\gamma})^2}^{(1+\sqrt{\gamma})^2} \frac{1}{x^2} dF_\gamma(x) = \frac{1}{(1-\gamma)^3} \tag{4.18}$$

This result is also confirmed by Remark 2 of Pan and Zhou (2011). □

Convergence of Term 2.2

Lemma 4.6. *Using the assumptions of previous lemmas.*

$$\frac{\mu_1^\top S^{-1} \Sigma S^{-1} \mu_2}{n} \xrightarrow{a.s.} \frac{a_2}{1-\gamma} \tag{4.19}$$

where $a_2 = \lim_{n\to\infty, p\to\infty} \frac{\mu_1^\top \Sigma^{-1} \mu_2}{n}$, $a_2 \neq 0$.

Proof. The proof follows from the same logic as the numerator term

$$4\frac{\mu_1^\top S^{-1} \Sigma S^{-1} \mu_2}{\mu_1^\top \Sigma^{-1} \mu_2} = \frac{(\mu_1+\mu_2)^\top S^{-1} \Sigma S^{-1} (\mu_1+\mu_2)}{(\mu_1+\mu_2)^\top \Sigma^{-1} (\mu_1+\mu_2)} \frac{(\mu_1+\mu_2)^\top \Sigma^{-1} (\mu_1+\mu_2)}{\mu_1^\top \Sigma^{-1} \mu_2}$$
$$- \frac{(\mu_1-\mu_2)^\top S^{-1} \Sigma S^{-1} (\mu_1-\mu_2)}{(\mu_1-\mu_2)^\top \Sigma^{-1} (\mu_1-\mu_2)} \frac{(\mu_1-\mu_2)^\top \Sigma^{-1} (\mu_1-\mu_2)}{\mu_1^\top \Sigma^{-1} \mu_2}$$
$$\frac{\mu_1^\top S^{-1} \Sigma S^{-1} \mu_2}{\mu_1^\top \Sigma^{-1} \mu_2} \to \frac{n}{1-\gamma} \tag{4.20}$$

□

Convergence of Term 2.3

Using Pan and Zhou (2011)'s derivation of high dimensional Hotelling T^2 statistics distribution, we now derive the convergence of the term with random variables on both sides of the quadratic form.

Lemma 4.7. *Following previous assumptions:*

$$\begin{aligned}
\frac{\bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1}{n||(\bar{Y}_1+\bar{Y}_2)||_2^2} &\to 0 \\
\frac{\bar{Z}_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_2}{n||(\bar{Y}_1+\bar{Y}_2)||_2^2} &\to 0 \\
\frac{\bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_2}{n||(\bar{Y}_1+\bar{Y}_2)||_2^2} &\to 0
\end{aligned}$$

Proof. Using similar transformation as before and and writing $(\bar{Z}_1+\bar{Z}_2)^\top\Sigma^{-1}(\bar{Z}_1+\bar{Z}_2)$ as $||\Sigma^{-1/2}(\bar{Z}_1+\bar{Z}_2)^\top||_2^2$

$$\begin{aligned}
\frac{(\bar{Z}_1+\bar{Z}_2)^\top S^{-1}\Sigma S^{-1}(\bar{Z}_1+\bar{Z}_2)}{(\bar{Z}_1+\bar{Z}_2)^\top\Sigma^{-1}(\bar{Z}_1+\bar{Z}_2)} &= \frac{(\bar{Z}_1+\bar{Z}_2)^\top\Sigma^{-1/2}}{||(\bar{Z}_1+\bar{Z}_2)^\top\Sigma^{-1/2}||_2}\Big(\frac{1}{n}Y^\top HY\Big)^{-2}\frac{\Sigma^{-1/2}(\bar{Z}_1+\bar{Z}_2)}{||\Sigma^{-1/2}(\bar{Z}_1+\bar{Z}_2)||_2} \\
&= \frac{(\bar{Y}_1+\bar{Y}_2)^\top}{||(\bar{Y}_1+\bar{Y}_2)^\top||_2}\Big(\frac{1}{n}Y^\top HY\Big)^{-2}\frac{(\bar{Y}_1+\bar{Y}_2)}{||(\bar{Y}_1+\bar{Y}_2)||_2} \qquad (4.21)
\end{aligned}$$

We let $u=\bar{Y}_1+\bar{Y}_2$. Since H is low rank and $\left(\frac{1}{n}Y^\top Y\right)^{-2}$ is a function of the sample covariance matrix, we can apply Theorem 2 of Pan and Zhou (2011) .

$$\frac{u^\top\left(\frac{1}{n}Y^\top Y\right)^{-2}u}{||u||^2} \xrightarrow{a.s.} \int\frac{1}{x^2}dF_\gamma(x)=\frac{1}{(1-\gamma)^3} \qquad (4.22)$$

Where $F_\gamma(x)$ is the Marchenko-Pastur Law and the LSD of the spectral distribution of the sample covariance matrix $\frac{1}{n}Y^\top Y$. Note that $||u||_2^2$ is of constant order, we then apply scaling of n, the convergence then becomes:

$$\frac{u^\top\left(\frac{1}{n}Y^\top Y\right)^{-2}u}{n||u||^2} \xrightarrow{a.s.} 0 \qquad (4.23)$$

To prove the convergence of $\frac{\bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1}{(\bar{Z}_1+\bar{Z}_2)^\top\Sigma^{-1}(\bar{Z}_1+\bar{Z}_2)}$ and $\frac{\bar{Z}_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_2}{(\bar{Z}_1+\bar{Z}_2)^\top\Sigma^{-1}(\bar{Z}_1+\bar{Z}_2)}$, we can follow similar logic as lemma 3.1 of Bai et al. (2009) and show that it converges to 0. Thus, expanding the the above result, $\frac{2\bar{Z}_2^\top S^{-1}\Sigma S^{-1}\bar{Z}_2}{(\bar{Z}_1+\bar{Z}_2)^\top\Sigma^{-1}(\bar{Z}_1+\bar{Z}_2)}$ also converges to 0.

□

Convergence of term 2.4

$$\begin{aligned}\frac{\mu_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1}{n||\bar{Y}_1+\bar{Y}_2||_2} &\leq \sqrt{\frac{\mu_1^\top S^{-1}\Sigma S^{-1}\mu_1}{n}}\sqrt{\frac{\bar{Z}_1^\top S^{-1}\Sigma S^{-1}\bar{Z}_1}{n||\bar{Y}_1+\bar{Y}_2||_2^2}} \\ &\leq 0\end{aligned} \tag{4.24}$$

Similar logic applies to all other terms in 2.4.

Combining all terms

Pulling all the terms together and recall that we define the scaled mahalanobis distance as: $\Delta_*^2 = a_1 + a_2 - 2a_3 = \lim_{n\to\infty,p\to\infty}\frac{(\mu_1-\mu_2)^\top\Sigma^{-1}(\mu_1-\mu_2)}{n}$.

Non-zero terms in the numerator term for group $k = 1$:

$$\begin{gathered}\frac{1}{n}\Big(\frac{1}{2}\mu_1^\top S^{-1}\mu_1 - \mu_1^\top S^{-1}\mu_2 + \frac{1}{2}\mu_2 S^{-1}\mu_2\Big) \xrightarrow{a.s.} \\ \frac{\frac{1}{2}a_1 - a_3 + \frac{1}{2}a_2}{(1-\gamma)}\end{gathered} \tag{4.25}$$

Numerator term for group $k = 2$:

$$\begin{gathered}\frac{1}{n}(-\frac{1}{2}\mu_2^\top S^{-1}\mu_2 + \mu_2^\top S^{-1}\mu_1 - \frac{1}{2}\mu_1 S^{-1}\mu_1) \xrightarrow{a.s.} \\ \frac{-\frac{1}{2}a_1 + a_3 - \frac{1}{2}a_2}{(1-\gamma)}\end{gathered} \tag{4.26}$$

Combining the terms and multiplying by (-1) for the first group:

$$\frac{-\Delta_*^2}{(1-\gamma)}$$

For the denominator, we apply the scaling of $\frac{1}{n||\bar{Y}_1+\bar{Y}_2||_2^2}$. We also notice the limit of $||\bar{Y}_1+\bar{Y}_2||_2^2$ is 4γ (confirmed by Li and Yao (2016)). Thus, $\frac{\mu_k S^{-1}\Sigma S^{-1}\mu_k}{n||\bar{Y}_1+\bar{Y}_2||_2^2} \xrightarrow{a.s.} \frac{a_k}{4\gamma(1-\gamma)^3}$. All other terms with random vectors converge to zero, the denominator term then becomes:

$$\sqrt{\frac{\Delta_*^2}{4\gamma(1-\gamma)^3}} \tag{4.27}$$

Combining the terms, we get the desired result.

$$R_{\text{LDA}} \xrightarrow{a.s.} \Phi\Big\{-2\Delta_*\sqrt{\gamma(1-\gamma)}\Big\}$$

4.3 Discussion of result

4.3.1 Behaviour of Error Rate in Under-parameterized Regime

To interpret the result, we firstly notice that error is dependent on the scaled Mahalanobis distance Δ_* and γ. The error has a negative relationship with Δ_*: as Δ_* increases, the value inside the normal CDF Φ becomes more negative, making its value decreases from random guessing to 0 ($\Phi(0) = 0.5$ to $\Phi(-\infty) = 0$). We also notice that Δ_* is dependent on γ. From Figure 4.1, we notice that Δ_* also increases with γ, though the rate of increase is compromised if the covariance matrix deviates from identity.

On the other hand, when γ = 0 and $\gamma = 1$, we both get $\Phi(0) = 0.5$. we observe that there is parabolic relationship between γ and error rate, where minimum error is achieved at$\gamma = 0.5$. This means there is a peaking phenomenon at $p \simeq n$ if we fix Δ_*. The $\sqrt{\gamma(1-\gamma)}$ term is the extra term we obtain comparing to the Bayes error and represents the cost of estimation for not knowing population parameters (μ_1, μ_2, Σ), which first decreases as γ increase and rises again. From simulation in Figure 4.2, for the identity covariance matrix case we observe the classical U-shaped curve where error is minimum at $\gamma \in (0.4, 0.8)$, and rises sharply as $\gamma \to 1$. Introducing feature correlation raises the error rate and move the minimum point to smaller values of γ. This because the rate of increase in Δ_* is smaller for each γ. and error is quickly

dominated by the increase in estimation cost.

We note that this result is quite similar to the one observed by Wang and Jiang (2018) under gaussian assumption (See Appendix) and corresponds to our classical understanding of U-shaped error curve in the under-parameterized region. We summarize the shape of the error curve as the interplay between increase in discrimination power through Δ_* and difficulty in estimating the population parameters including conditioning of the covariance matrix inverse. We summarize the behaviour of the error curve as below.

Behaviour I $\gamma \in (0, 0.5)$: In this region, increasing the dimension relative to the sample size decreases the classification error from random guessing because of the increase in discriminatory information. The local minima occur at $\gamma \approx n/2$ similar to that observed by Raudys and Duin (1998).

Behaviour II $\gamma \in (0.5, 1)$: Increasing dimension relative to sample size increases model error as the estimation cost increases. As $\gamma \approx 1$, the sample covariance matrix is ill-conditioned, and the estimation is highly unstable, leading to the peak in error.

Note on the Normal Assumption

The contribution of our result is that we used existing RMT literature to derive the convergence of LDA's error rate without the normal assumption. However, since Φ is used which means that the entire term is still enveloped within the normal CDF. For future work, we aim to find the limit of the difference between an arbitrary cumulative distribution function and the normal CDF. It may be possible to prove that the difference converges to 0 asymptotically. This would make the result more complete. However, since our result differs from those derived under normal assumption, it is more general and provides additional insight.

4.3.2 Behaviour of Error Rate in Over-parameterized Regime

It is more difficult to to derive the result in the over-parameterized regime where $\gamma > 1$ and the Moore-Penrose pseudo-inverse is used. The standard M-P law applies less easily to the $\gamma < 1$ case. We could make restrictive assumptions such as assuming that the population covariance

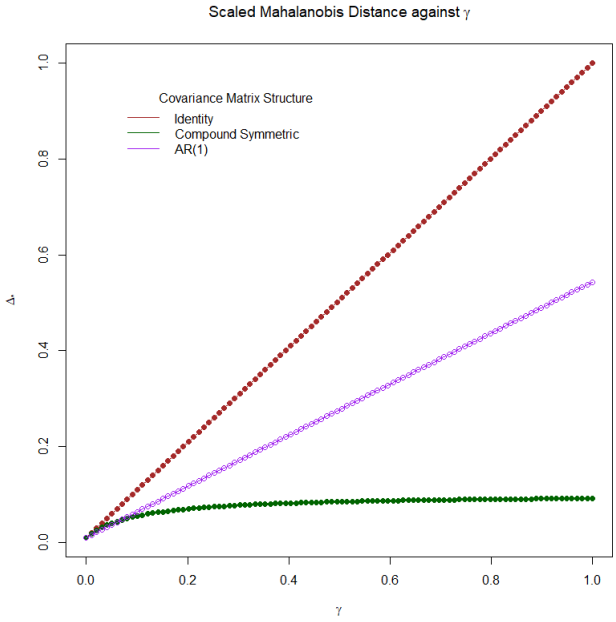
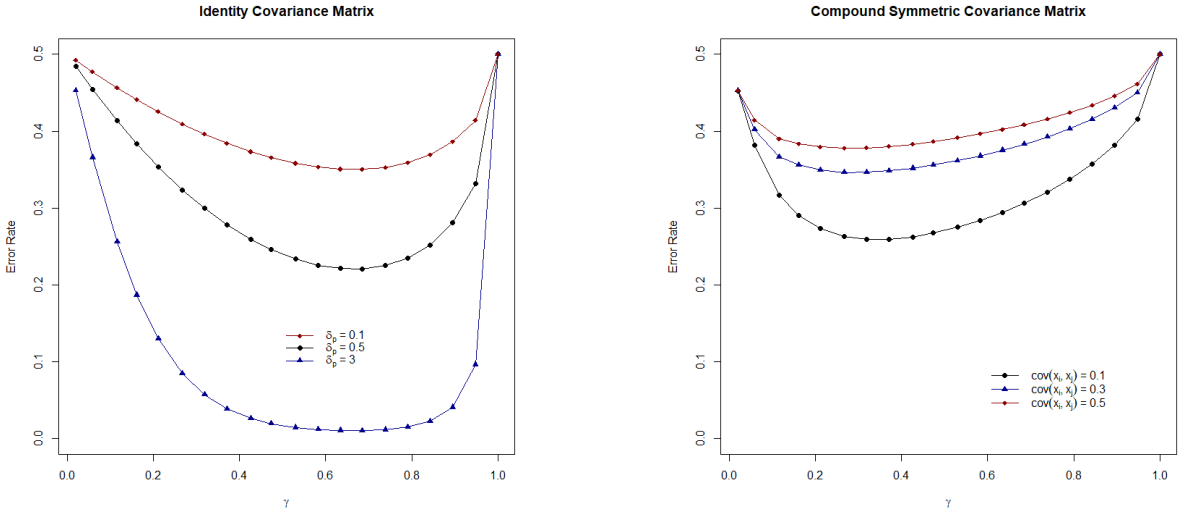


Figure 4.1: Scaled Mahalanobis distance Δ_* against γ for different covariance matrix structure ($\mu_1 - \mu_2 = 1_p, n = 100$; AR(1) $\rho = 0.3$, Compound Symmetric with covariances 0.1).



(a) Identity Covariance Matrix and different population mean separation $\delta_p = \mu_1 - \mu_2$

(b) Compound symmetric covariance matrix $\delta_p = 3$

Figure 4.2: Visualization of Theorem 1 ($n \in (20, 400), \gamma \in (0, 1)$)

matrix is multiple of the identity $\Sigma = \sigma I$ to overcome certain restrictions. However, one potential approach is by taking the limit of the error convergence of the regularized discriminant analysis (RLDA). In the supplementary paper of Hastie et al. (2022), the authors take the limit of a ridge regularized term to find the bias term of the minimum norm least square regression

by noting the following characterization of the pseudo-inverse of a rectangular matrix A:

$$(A^\top A)^+ A^\top = \lim_{z \to 0^+} (A^\top A + zI)^{-1} A^\top$$

we also note that the error expression of the RLDA is:

$$R_{RLDA}(\lambda) = \Phi\Big(\frac{(-1)^k \Big(\mu_k - \frac{\bar{X}_1 + \bar{X}_2}{2}\Big)^\top (S + \lambda I_p)^{-1} (\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^\top (S + \lambda I_p)^{-1} \Sigma (S + \lambda I_p)^{-1} (\bar{X}_1 - \bar{X}_2)}} \Big) \quad k = 1, 2 \tag{4.28}$$

Leveraging Theorem 3.1 of Wang and Jiang (2018), we deduce the preliminary result

Lemma 4.8. *Assuming the following conditions*

1. $p/n \to \gamma \in (1, \infty)$
2. *The eigenvalues of Σ are uniformly bounded and the ESD of F^Σ converges to a nonrandom distribution function H as $p \to \infty$*
3. *For $t \geq 0$:*

$$\Delta^{-2} \mu^\top (I_p + t\Sigma^{-1})^{-1} \mu \to h_1(t)$$
$$\Delta^{-2} \mu^\top (I_p + t\Sigma^{-1})^{-2} \mu \to h_2(t)$$

$\mu = \Sigma^{-1/2}(\mu_1 - \mu_2)$, $\Delta = \sqrt{(\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2)}$. Then the error of pseudo-inverse LDA converges in probability to

$$R_{PLDA} \to \Phi\Big(-\frac{H_1 \Delta^2}{2\sqrt{H_2 \Delta^2 + 4\gamma R_2}} \Big) \tag{4.29}$$

Where

$$
\begin{aligned}
R_1 &= \lim_{\lambda\to 0}\Big\{\frac{1-\lambda m_0(-\lambda)}{\{1-\gamma[1-\lambda m_0(-\lambda)]\}^3} - \frac{\lambda m_0(-\lambda)-\lambda^2 m_0'(-\lambda)}{\{1-\gamma[1-\lambda m_0(-\lambda)]\}^4}\Big\}\\
R_2 &= \lim_{\lambda\to 0}\Big\{\frac{1-\lambda m_0(-\lambda)}{1-\gamma[1-\lambda m_0(-\lambda)]}\Big\}\\
H_1 &= \lim_{\lambda\to 0}\Big\{\frac{1}{1-\gamma[1-\lambda m_0(-\lambda)]}h_1\Big(\frac{\lambda}{1-\gamma[1-\lambda m_0(-\lambda)]}\Big)\Big\}\\
H_2 &= \lim_{\lambda\to 0}\Big\{\{(1+\gamma R_2(\lambda))^2+\gamma R_1(\lambda)\}h_2\Big(\frac{\lambda}{1-\gamma[1-\lambda m_0(-\lambda)]}\Big)\Big\}
\end{aligned}
$$

We require that the limits r_1, R_1, H_1, H_2 exist. $m_0(-\lambda)$ is the unique solution of the Marchenko-Pastur equation:

$$
m(-\lambda) = \int \frac{dH(t)}{t(1-\gamma+\gamma\lambda m(-\lambda))+\lambda}
$$

Under the condition $1-\gamma+\gamma\lambda m(-\lambda) \geq 0$. If we are working in the under-parameterized regime $\gamma < 1$, we can applying the Stieltjes transform and Lemmas 7 and 8 of Cheng et al. (2022) , we would obtain the result$\lim_{\lambda\to 0} m_0(-\lambda) = \frac{1}{1-\gamma}$, $\lim_{\lambda\to 0} m_0(-\lambda) = \frac{1}{(1-\gamma)^3}$ and $R_1 = \frac{1}{1-\gamma}$, $R_2 = \frac{1}{(1-\gamma)^3}$. However, the behaviour of $\lim_{\lambda\to 0} m_0(-\lambda)$ is not as straightforward in the over-parameterized regime and further work is required to derive the limit. We could consider applying result of Ledoit and Péché (2011), which derived the asymptoptics of the functional of the type $\frac{1}{n}\text{tr}(g(\Sigma)(S-\lambda I)^{-1})$ in the setting of $\gamma > 0$ and certain moments conditions. We leave this problem as suggestion for future work.

Hypothesized Behaviour III $\gamma > 1$ when $\gamma > 1$, Bai and Yin (1993) noticed that the condition number of the sample covariance matrix is proportional to $(1-\sqrt{1/\gamma})^{-2}$ and decreases with increasing γ. Similarly, Hoyle (2010) provided us insight on the behaviour of pseudo-inverse in high dimensions through suggesting that the difference with true inverse decreases with increasing γ.

From Cheng et al. (2022), LDA's error convergence in $\gamma > 1$ case where $n_1 = n_2$ when using pseudo-inverse and assuming Gaussian features is found to only differ by an extra γ term in the denominator and the $\sqrt{1-\gamma}$ term in the numerator changes to $\sqrt{\gamma-1}$.

From these information, we hypothesize that the error curve will first descent from 0.5. It will reach a minimum point when γ is a small constant value. However, the rate of increase of the error and the second descent will depend on the covariance matrix structure. The second descent may persist for a longtime for identity population covariance matrix, as its eigenvalues are constant. For other covariance matrix structure with exploding or fastly decaying eigenvalues, the error will likely eventually rise.

Connection to Benign Overfitting

We recall that benign overfitting refers to the phenomenon where increasing features to close to or more than the sample size leads to improvement in error rate (Bartlett et al., 2020) . Our analysis of the error rate in the underparameterized regime partially answers the question for LDA by observing that there is a peaking phenomenom and providing a theoretical explanation for it as in linear regression. We also notice that in the double asymptotic setting, the relationship between LDA error and model flexibility is largely influenced by the dimension effect of estimating the population parameters. In the next section, we will further investigate this phenomenon using simulated data, with focus on the $\gamma > 1$ regime.

5 Simulation Analysis

5.1 Aim

In this chapter, we conduct a simulation study to verify whether LDA's classification error exhibits double descent at different values of γ. The goal of this simulation analysis is to investigate the shape of the error curve under different scenarios which approximate real-life data: for example, effect of changing sample size, different population covariance matrix Σ structure, data that violate normality assumption, noisy observations and redundant observations. Since our theoretical work concerns with the under-parameterized regime, we complement it by considering the $\gamma > 1$ to see whether the double descent curve occurs. For the different simulation settings, we are interested in describing the location of the local and global minimum error to understand the benign over-fitting and inform real-life practices of setting the optimal over-parameterization ratio γ.

Specifically, we aim to answer the following research questions:

Q1. Effect of sample size: Is increasing sample size always beneficial for model performance?

Q2. Effect of different covariance matrix structure: How would different covariance matrix structures, such as incorporating different feature correlations, affect LDA error rate compared to the identity case?

Q3. Effect of non-normality: Does the double descent shape hold for data that is non-normal?

Q4. Effect of noisy observations: Would over-parameterization help model performance in the case of noisy data or mislabelled response?

Q5. Effect of Redundant features: Does non-discriminatory features change the shape of the error and its peak?

5.2 Data Generating Process

Even though our theory analysis is under the double asymptotic setting $p \to \infty, n \to \infty$, we will not attempt to approximate this setting in the simulation as the computation of the generalized inverse with large p is time-consuming. Since previous simulations have shown that asymptotic results can still provide good guidance of finite sample analysis (Cheng et al., 2022), we will use a small value of n.

The main data-generating process is as follows:

Step 1 Create a sequence of sample size for each group where group sample size n_{1j}, n_{2j} $(j = 1, \ldots, 20)$ varies from 20 to 50 and dimension p_j from < 5 to 400 so that their ratio $p_j/(2n_{kj}) = \gamma_j$ varies from 0.5 to 4. (The training data set is balanced for each group n_{kj}).

Step 2 Determine population parameters μ_1, μ_2, Σ. (In most cases, we let the mean vector have the same entries and control their difference $\mu_1 - \mu_2 = \delta_p$). (The dimension of the parameter varies based on p_j).

Step 3 Loop through the index j. At each γ_j, generate data from multivariate normal distribution with sample size n_j and dimension p_j: $X_{1i} \sim \text{MVN}_{p_j}(\mu_{1p_j}, \Sigma_{p_j}), X_{2i} \sim \text{MVN}_{p_j}(\mu_{2p_j}, \Sigma_{p_j})$, $i = 1, \ldots, n_j$.

Step 4 Let response variable of each group be $Y_1 = -1$, $Y_2 = +1$.

Step 5 Combine the data and split into 70 percent training n_{train} and 30 percent testing n_{test}.

Step 6 Construct linear discriminant rule with pseudo-inverse on the training data and use the rule to obtain prediction on testing data $\widehat{Y}_i, i = 1, \ldots n_{test}$.

Step 7 Calculate error rate: $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (Y_i \neq \widehat{Y}_i)$.

Step 8 For each j, repeat steps 3-7 100 times and average the error rate.

To confirm the empirical simulation result, we also apply the theoretical conditional error. we employ the same data-generating process. Rather than predicting on a testing set using sample

estimations $\bar{X}_k, \widehat{\Sigma}$, we calculate the error rate using the below formula for R_{LDA}. Results are also averaged over 100 trials.

$$R_{\text{LDA}} = \frac{1}{2}\sum_{k=1}^{2}\Phi\left\{\frac{(-1)^k\{\mu_k - \frac{1}{2}(\bar{X}_1 + \bar{X}_2)\}^\top S^{-1}(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\bar{X}_1 - \bar{X}_2)^\top S^{-1}\Sigma S^{-1}(\bar{X}_1 - \bar{X}_2)}}\right\} \quad (k = 1, 2) \tag{5.1}$$

5.3 Simulation Results

5.3.1 Effect of sample size

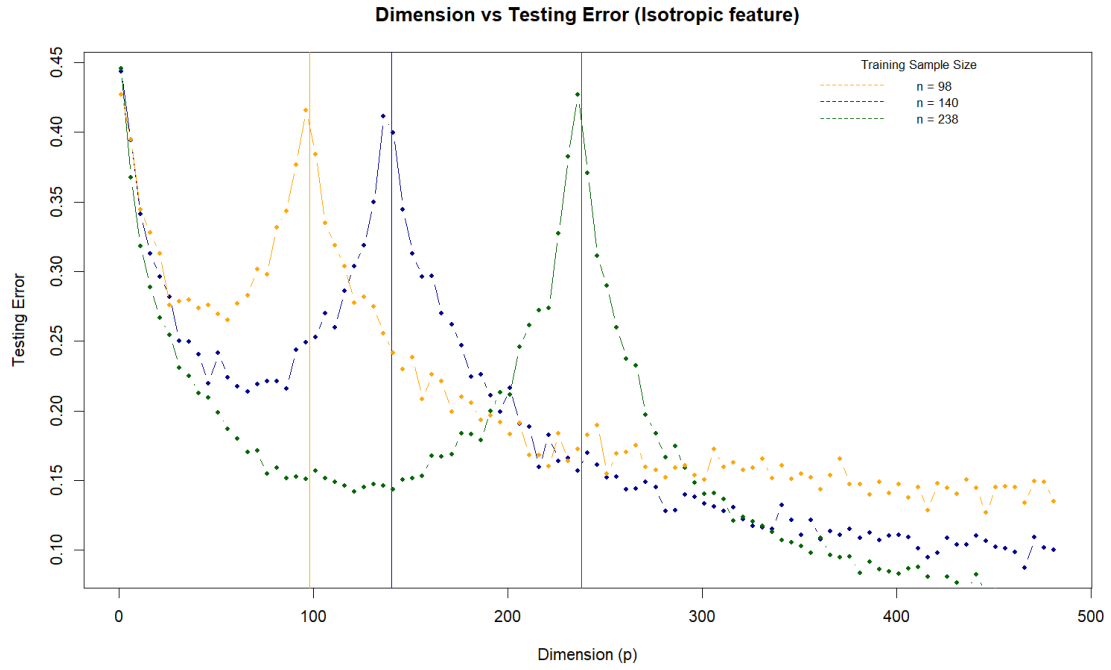


Figure 5.1: Simulation error rate of identity covariance matrix for fixed n and increasing p ($\delta_p = 0.1, \Sigma = I_p$)

To investigate the effect of size in a high dimensional setting, we fix the sample size and increase dimension, and the error curves exhibit double descent, showing a U-shape in the under-parameterized regime, peaks at $p \simeq n$ and continuously decrease in over-parameterized regime. We notice an interesting phenomenon that at different values of p, more training sample does not always lead to better performance. For example, when $p = 150$, sample size $n = 98$ performs better than $n = 140$, since the smaller sample error curve has reached the second

descent and the larger sample curve has only started to decrease from the peak. Hence, we conclude that larger sample sizes could hurt model performance depending on their ratio γ. The sample size should either be around twice as less as p or twice as more for the model to have good performance. This observation is in accordance with the behaviour noticed by Deng et al. (2021) for logistic regression with polynomial features: the double descent curve for fixed n, increasing p also overlap each other, suggesting that more sample size could decrease predictive power. Similarly, the effect of sample size on neural network model performance is studied empirically (Nakkiran et al., 2021): increasing sample size has the effect of shifting the interpolation threshold to the right leading to a higher error rate.

5.3.2 Effect of varying covariance matrix structure

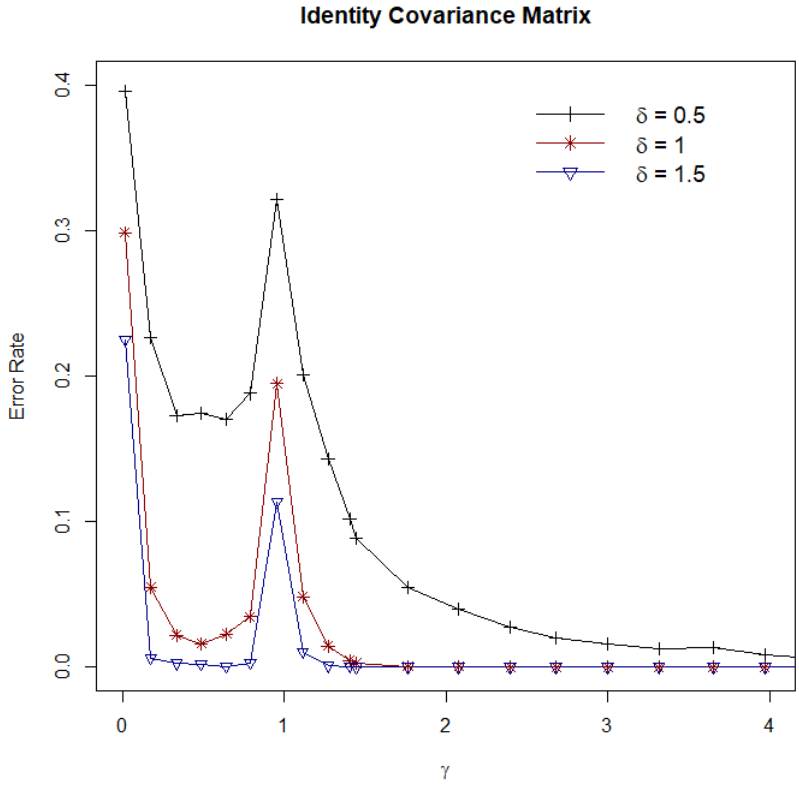


Figure 5.2: Simulated error rate of identity covariance matrix with increasing n, p, γ, with different mean separation $\delta_p \in (0.5, 1, 1.5)$

Isotropic covariance Structure $\Sigma = I_p$

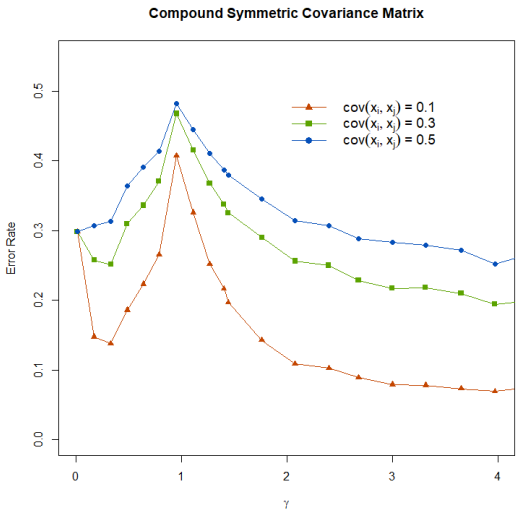
To consider the effect of covariance matrix structure, we first consider the case where the true covariance matrix is an identity as a baseline.In Figure 5.2, we increase both n and p and the test error becomes monotonically greater as δ_p decreases, corresponding to increasing difficulty in separating the classes. In the over-parameterized regime, the test error decreases to near 0 quite fast and does not rise for increasing γ. As expected, the cost of over and under-estimation or eigenvalues is small in the identity case as all eigenvalues are 1. As γ increases, the pseudo-inverse becomes better conditioned, which helps the model achieve near 0 error at a small over-parameterization.

Compound Symmetry Structure

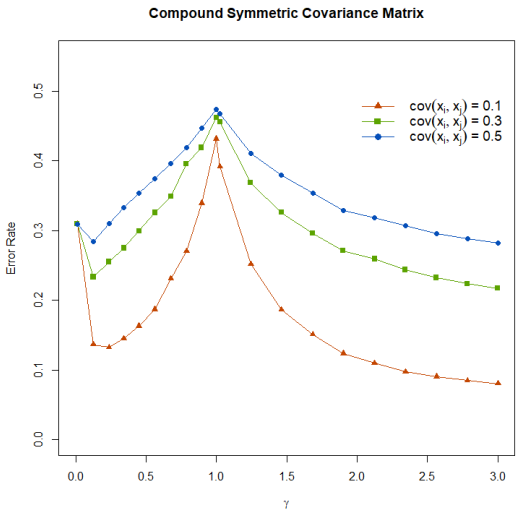
In empirical applications, the covariance matrix structure is likely more complex than the identity case. We introduce feature correlation in the compound symmetric structure. We again let the variance be 1 and assume that features all have the same covariance structure controlled by ρ.

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \rho & 1 \end{bmatrix}$$

As confirmed by both simulated data and theory in Figure 5.3 (a) and (b), the error rate is higher monotonically for higher σ as classification is more difficult when covariance increases. Unlike the identity case, the minimal error in the classical region no longer occurs at $\gamma = 0.5$, but rises quickly. This can be interpreted from the fact that the Mahalanobis distance rises more slowly when covariances exist, making it more difficult to estimate the true covariance matrix and separate the classes. In the $\gamma > 1$ region, the error decreases much slower than the identity case as it does not quickly go to 0, and the global minimum seems to occur at much larger γ (> 3). This is due to inexact estimation of the eigenvenalues (especially the largest one



(a) Simulation Error Rate



(b) Theoretical Error Rate

Figure 5.3: Compound symmetric covariance matrix ($\delta_p = 1_p$)

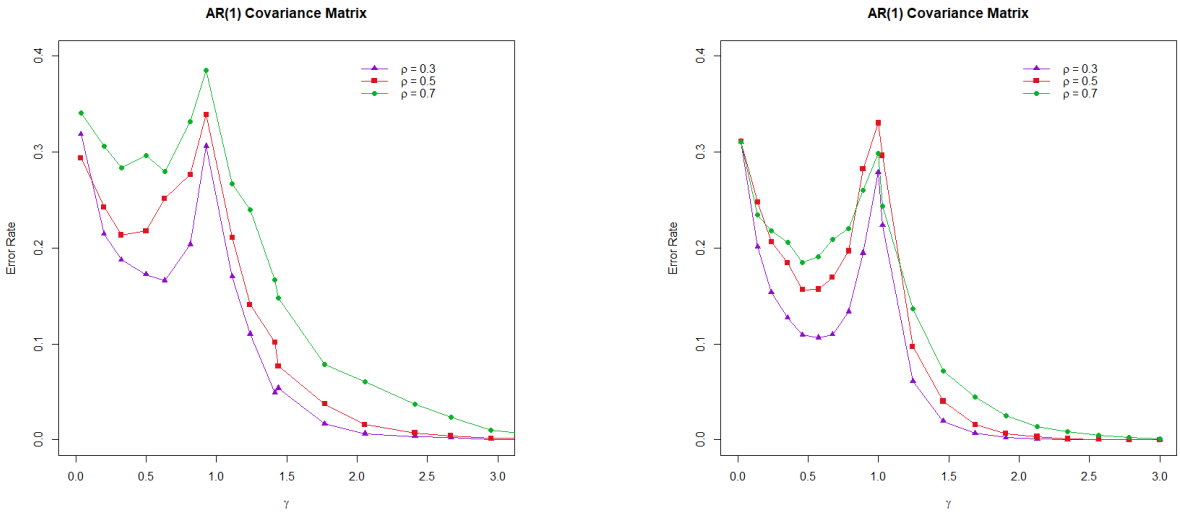
which grows linearly with p), the implicit regularization of the pseudo-inverse takes longer to overcome this effect.

First Order Autoregressive - AR(1)

The performance of LDA under AR(1) or first-order autoregressive covariance matrix structure is also worth examining, as this structure occurs in various real-life scenarios where successive observations are correlated. The strength of correlation decreases exponentially as lag increases. Its application includes changes in price in finance, economic indicators such as GDP growth, and environmental record readings.

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-1} \\ \rho & 1 & \rho & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \rho \\ \rho^{p-1} & \rho^{p-2} & \dots & \rho & 1 \end{bmatrix}$$

From examining simulated and theoretical result in 5.4 (a) and (b), the AR(1) covariance matrix structure behaves similarly to the compound symmetric case in the under-parameterized



(a) Simulation Error Rate

(b) Theoretical Error Rate

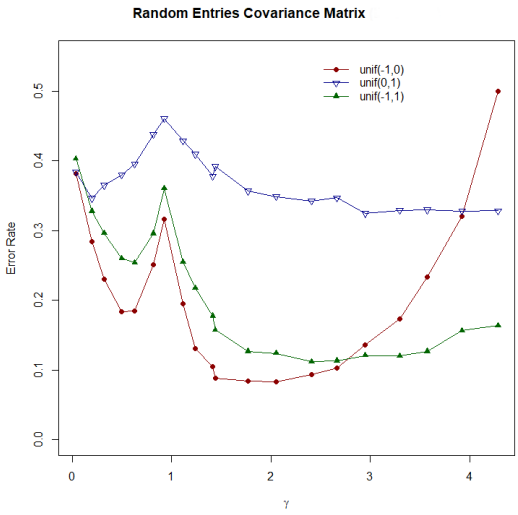
Figure 5.4: AR(1) covariance matrix structure with varying covariance structures, $\delta = 1_p$

regime as high covariances have monotonically higher error rates. However, as γ increases the difference between curves of different ρ is very small due to the reduced correlation strength. Therefore, the global minimum occurs also at the over-parameterized region and is much lower than the local minimum at under-parameterized region. This is because of the decay of eigvenalues when γ increases, causing it to behave like the identity matrix. Thus, when the covariance matrix exhibits AR(1) structure, over-parameterization could greatly improve performance.

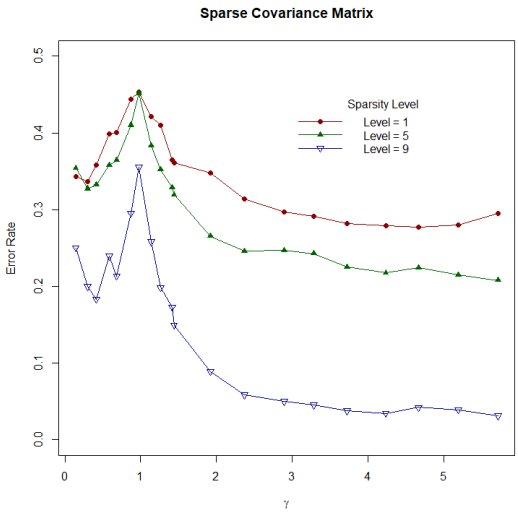
Random Entries and Sparse covariance matrix

We now consider an unstructured covariance matrix with diagonal elements around 1 and off-diagonal elements generated from a uniform distribution with different ranges (ensuring semi-positive definiteness by adding a small constant to the eigenvalues and reconstructing the matrix). This type of covariance matrix is more common in real-life datasets. In high dimensional setting, not all features are necessarily correlated with each other, so we also consider the case of sparse covariance matrix where some off-diagonal elements are set to close to 0. For all ranges of covariances, the peak in error rate occurs at $\gamma \approx 1$ and the global minimum is reached in the overparameterized cases. However, their shapes are slightly different. When

the range of covariance is between $[0, 1]$, the error rate is the highest out of the three cases but tails off in large γ. When the range is negative $[-1, 0]$, the error is the lowest when $\gamma < 1$ but rises quickly when to random guessing (0.5) as γ increases beyond 1. This is due to the multicollinearity issue when p increases as predictors become linearly dependent. When the range is $[-1, 1]$, the error rate combines the characteristics of the two other covariance matrices and has the tendency to slightly trend upwards at large γ. This could again be due to the interplay of covariance matrix conditioning and dominance of large eigevenvalues. Thus, in real-life situations, there will likely be a second increase in test error when features have positive and negative correlation.The global minimum would be moderate over-parameterization $\gamma \in (1, 3)$. When the matrix has high level of sparsity (many covariance values close to 0), we would prefer large overaparmeterization as it could achieve similar to the identity case. However, since the sparseness is random and unrelated to γ, curves will not converge together at large γ like the AR(1) case.



(a) Covariance matrix generated from uniform distribution of different ranges $X_{ki} \sim U[\min_k, \max_k]$, $k = 1, 2$

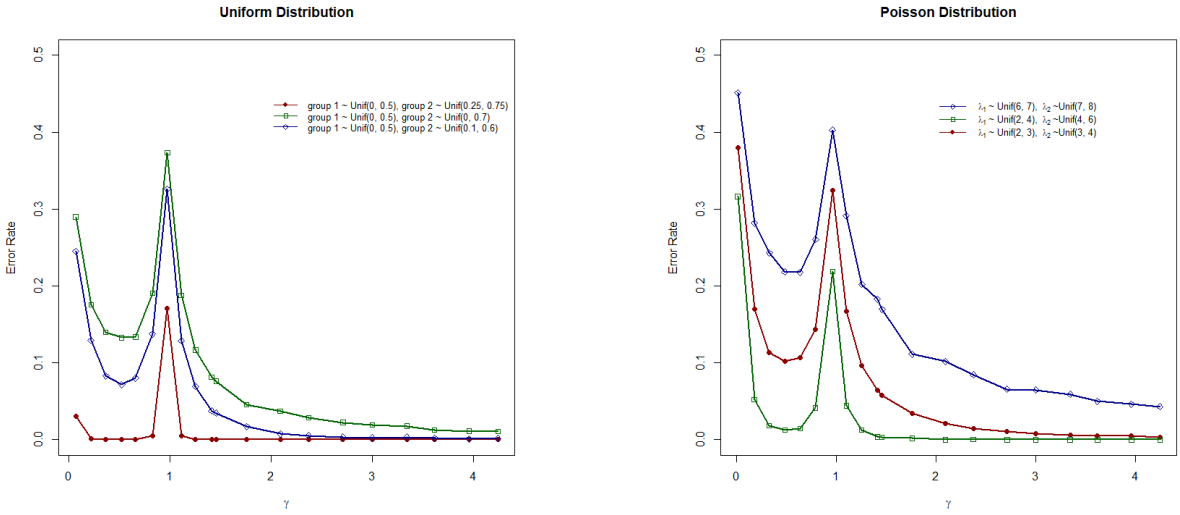


(b) Covariance Matrix with off-diagonal entries generated from $U[-1, 1]$, with a certain level set to near 0.

Figure 5.5: Random Entries Covariance Matrix ($\delta_p = 1_p$)

5.3.3 Non-normal Data

The true distribution of the data is often unknown in reality and LDA can be applied to datasets that follow a non-normal distribution. Since our novel theoretical contribution is using a proof technique that does not require X to be normally distributed, we will investigate the effects of violating the assumption of normality on the double descent phenomenon.



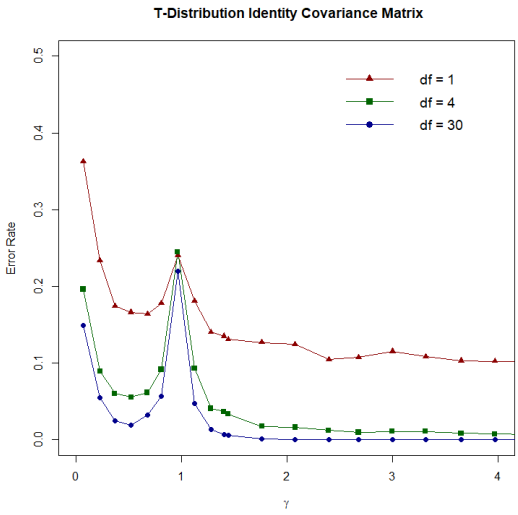
(a) Data Generated from uniform distribution $X_{ki} \sim U[\min_k, \max_k], k \in \{1, 2\}$

(b) Data Generated from Poisson distribution $X_{ki} \sim \text{pois}(\lambda_k), k \in \{1, 2\}$

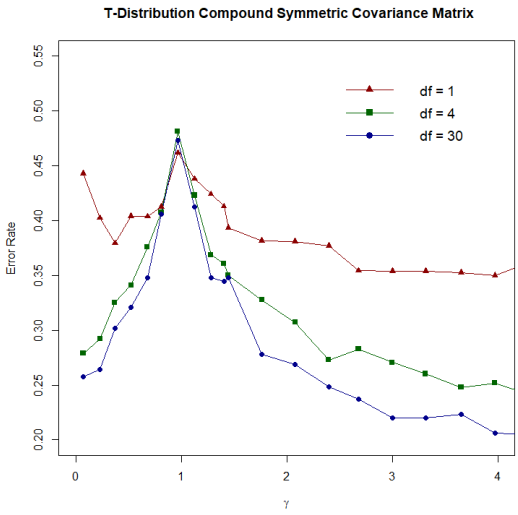
Figure 5.6: Independently distributed Features

We firstly consider the case where features, or columns of the data matrix, each follow independently uniform and Poisson distribution (Figure 5.6). We observe that LDA behaves similarly to the identity covariance matrix case. The test error peaks around $p \approx n$ and tails off asymptotically towards 0 for large γ. Thus, the benign overfitting still exists.

T-distribution has a heavier tail than normal distribution to allow for more extreme values. T-distribution also does not have finite fourth moment, which is a requirement of the theoretical result. However, simulation shows that when the degrees of freedom is < 4, double descent still exists but the error rate performance is poor and the benefit of over-parameterzation is largely compromised (Figure 5.7). The error rate in the over-parameterized regime does not descend continuously and is not much lower than the under-parameterized regime. With degrees of freedom is set to > 4, LDA still has good performance. In both cases of covariance matrix



(a) T-distribution with identity covariance matrix



(b) T-distribution with compound symmetric covariance matrix ($\rho = 0.3$)

Figure 5.7: T-distributed data

structure, the behaviour is very similar to the normal case. In the case of log-normal distributed data or Dirichlet distribution, the test error does not dependent on γ and fluctuates randomly. Double descent is not observed and performance is similar to random guessing. Hence, when the data is heavily non-normal, LDA should not be applied as mean and covariance matrix estimation is futile. Other flexible extensions to LDA should be applied instead.

5.3.4 Noisy Observations

We now consider the case where data is corrupted by noise, as is often the case with real observations. Firstly, we flip a percentage of response Y to the opposite sign, this makes some data have spurious relationship with the response. In the second case, Gaussian noise is added to a percentage of data columns, which increases their variance.

From Figure 5.8, we notice that flipping response raises overall error rate but increasing γ helps to continuously decrease error rate and having more dimensions still helps to reduce variance and overcome the effect of noise. From Figure 5.9, adding varying levels of gaussian noise to a fixed percentage of data has the opposite effect. Higher level of noise makes the test error rise faster as γ increases, exhibiting a second rise in error and the global minimum

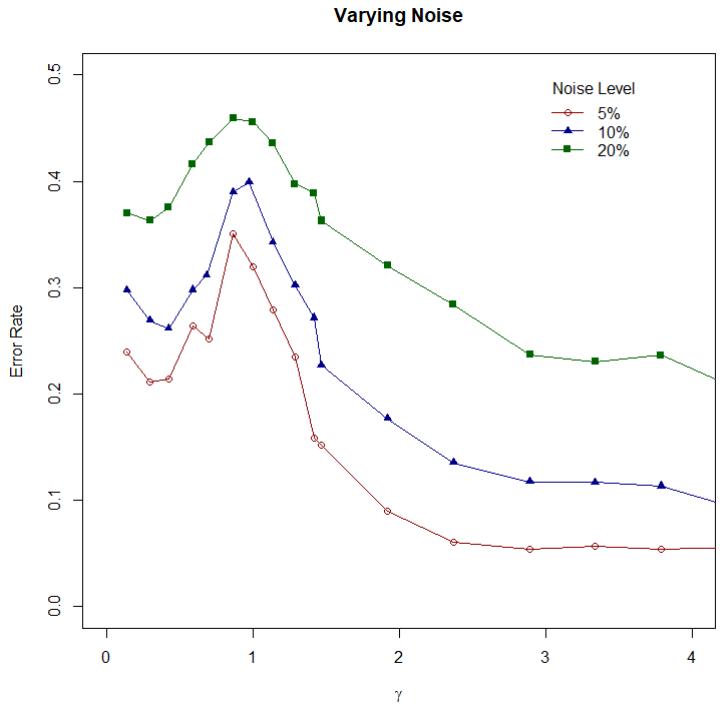


Figure 5.8: Flipping a percentage of response Y; $\delta_p = 1_p$, $\Sigma = I_p$

error occurs at small over-parameterization ratio $\gamma \in [1, 3]$. This suggests that one has to be careful about scaling features in practice if the data tends to have large variance, such as due to measurement error, as the minimum test error would only occur at a certain point of over-parameterization.

5.3.5 Redundant Features

To mimic real-life high dimensional data, we also set a percentage of data columns to be 0 so that they provide no discriminatory information to the decision rule. The result (Figure 5.10) shows that the peak in error no longer occurs around $\gamma \simeq 1$, but at larger values if redundant feature increases. This can be explained by the fact that "effective" number of p is smaller to achieve $p \simeq n$. However, if more redundant feature exists the error rate is higher at larger γ values (in over-parameterized regime). This observation is in accordance with that noticed by Chatterji and Long (2021): when irrelevant variables increase, the performance of the max-margin classifier degrades with γ.

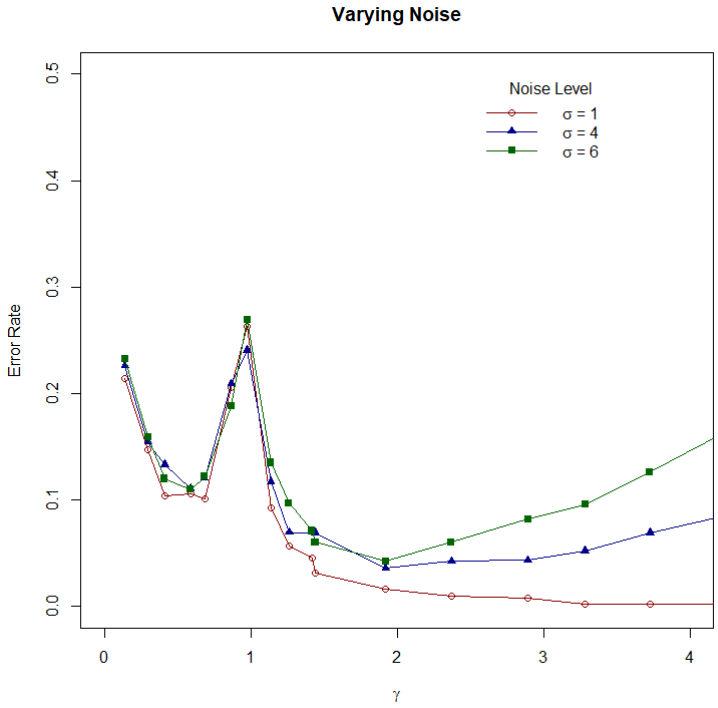


Figure 5.9: Adding Gaussian noise to 15 percent of data: $X_{ki} = X_{ki} + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$;$\delta_p = 1_p$, $\Sigma = I_p$

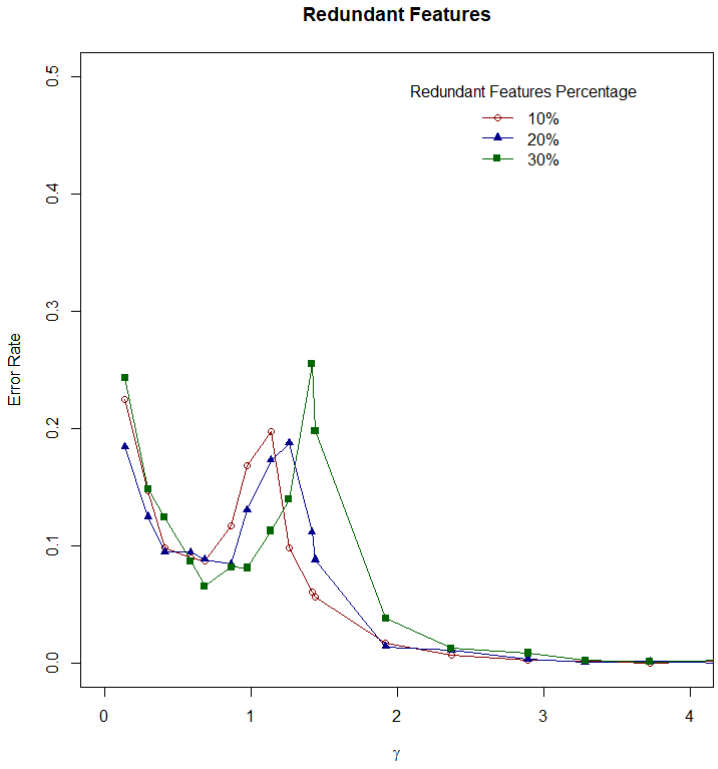


Figure 5.10: Varying the percentage of features that are non-discriminatory ($\delta = 1_p, \Sigma = I_p$)

5.4 Conclusion and Limitation

Overall, we were able to verify the hypothesis that double descent phenomenon is observed under various settings of Σ, mild model mis-specification and noise. In all cases, the global minimum is attained at the over-parameterized region. In certain cases such as unstructured covariance matrix, the error is much lower than under-parameterized cases. However, the behaviour of the error curve in $\gamma > 1$ setting varies and can ascend quickly for small γ values. Hence, large over-parameterization is likely not ideal in most real data scenarios. The key results can be summarized as below.

Q1 Effect of Sample size: Increasing sample size is beneficial for error rate at high values of over-parameterization ratio γ, but not necessarily in the under-parameterized regime when γ is close to 1.

Q2 Effect of different covariance matrix structure: Reduced feature correlation in high dimensions, such as with AR(1), Sparse covariance matrix structure, amplifies the benefit of overfitting. However, multicollinearity in high dimensions may cause a second rise in error.

Q3 Effect of non-normality: Double descent hold for independently distributed data and mild violation of normality but not for heavily skewed distributions.

Q4 Effect of noisy observations: If a certain percentage of data is corrupted by noise, increasing over-parameterization hurt model performance. However, if there is spurious relationship with the response, over-parameterization helps performance.

Q5 Effect of Redundant Features: Non-discriminatory features increases the value of γ where peaking occurs and larger γ value is needed to achieve the global minimum.

Overall, in real practice, we should be mindful of not to train an LDA model with similar number of n and p but to either reduce sample size or include more features to avoid the peak on error. A dataset with high feature correlation requires large over-parameterization to achieve low error. Hence, over-parameterization is worth considering but many factors at play influences the value of γ at which the global minimum error will occur. As a general rule, increasing the features to 2 times the training sample size will lead to lower error than the

under-parameterized regime. However, removing redundant feature and checking for linearly dependent features is necessary. Ignoring off-diagonal elements in the covariance matrix has also been proven to have good performance on unseen datasets (Dudoit et al., 2002).

6 Empirical Data Analysis

6.1 Aim

In previous chapters, we have focused on discussing the error rate from a theoretical point of view and from synthetic data. In practice, the true population parameters of the data are unknown, the observations can be noisy and the data may violate various LDA's assumptions. It is important to check if double descent still occurs in a real-life scenario. Hence, we will perform real data analysis on the ARCENE cancer classification dataset (Guyon et al., 2008). We hypothesize that the double descent curve will appear as suggested by theory, simulation, and previous literature.

We will also compare the performance of pseudo-inverse LDA with other high-dimensional LDA methods that existing literature has developed. This helps to understand the competitiveness of the pseudo-inverse method. Since the goal of studying double descent is to uncover whether high model flexibility can lead to a low error rate, we should investigate whether over-parameterization can make the pseudo-inverse LDA (PLDA)'s error rate smaller than other methods. We will consider dimension reduction using PCA, diagonal discriminant analysis (DLDA), shrunken centroid discriminant analysis (SLDA) and maximum uncertainty LDA (MLDA). The result should shed light on whether and when we should consider the pseudo-inverse method in real practice.

6.2 Background and Data Exploration

The goal of the classification is to distinguish cancer versus normal patterns from mass-spectrometric data. Mass spectrometry is a technique that identifies and quantifies the chemical composition by measuring the mass-to-charge ratio of molecules. The features are the level of proteins in human Sera for a given mass value. The response takes 2 values, $y = +1$ indicates cancer patients and $y = -1$ for healthy patients. The original training set consists of $p = 10,000$ features with $n = 100$ instances of randomized order and pattern. 3000 of the features are probe features with no discriminatory power. All features are numerical and continuous.

Model Checking

Model checking is necessary before conducting LDA as simulation has suggested that its performance is poor when model assumptions are heavily violated. However, due to the large number of features compared to the sample size, it is hard to validate assumptions for all features; many hypothesis tests also do not work in high-dimensional settings. Nevertheless, we randomly select a subset of features to perform preliminary model checks to understand the data.

Normality Assumption: Recall that one of the assumptions of LDA is that the data is multivariate normally distributed, which means that each variable follows univariate distribution. Visually inspecting the Q-Q plot of randomly selected samples, normality is heavily violated for most variables due to skewed distribution and many observations being 0. Selectively conducting Mardia's skewness and kurtosis test (MARDIA, 1970) on a random subset of features also show that the majority of groups of variables fail the multivariate normality assumption.

Equality of covariance matrix structure: Examining the covariance matrix, most variables have very large variance ($1000+$) and a large range of non-zero covariances (from < 1 to $1000+$). Conducting Box's M test on a subset of features also shows p-value < 0.05 for almost all subsets, which suggests that the null hypothesis that covariances are equal across groups should be rejected.

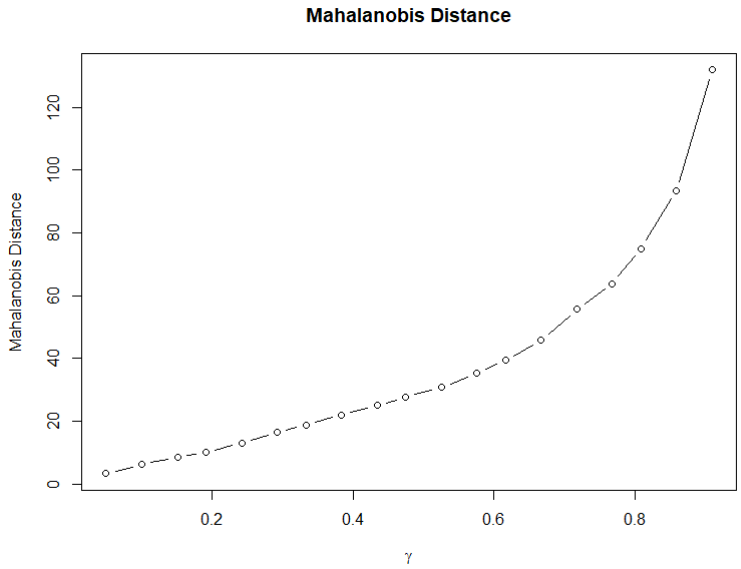


Figure 6.1: Mahalanobis Distance for $\gamma \in (0, 1)$ (Result is averaged over 50 trials of randomly selected features)

Mahalanobis distance vs feature: We have learned from theory that the discriminatory ability depends on the true Mahalanobis distance of the two groups, which influences the shape of the error curve. Therefore, we also check how the distance of this dataset vary with changing γ. From the graph of randomly selected features' averaged Mahalanobis distance, it is evident that the distance between the two groups increases with γ. The exponential increase in the distance suggests that increasing features relative to the training sample may help with decreasing the error rate, especially at large over-parameterization.

Hypothesis

Overall, the dataset violates the normality and equal covariance assumption, but since we know that double descent is robust in case of certain model assumption violation and the Mahalanobis distance varies with γ, we still expect double descent to appear when applying PLDA and the minimum error should occur in the $\gamma > 1$ region.

6.3 Methodology

To compute the error rate, we randomize the order of the features and then vary gamma from 0 to 10. The error rate is calculated from $\frac{1}{n_t}\sum_{i=1}^{n_t} \mathbb{1}(Y_i \neq \widehat{Y}_i)$ using a testing set of $n_t = 99$ and the process is repeated 30 times to get an average estimate.

We will also consider applying dimension reduction through principal component analysis. PCA works by finding the eigen-decomposition of the covariance matrix to identify orthogonal directions of the data where variance is maximized. We will fit the LDA model with the varying number of principal components (since the rank of the covariance matrix is at most $n - 1$ there will only be 98 non-trivial components).

Various techniques have been developed to handle LDA in the $n < p$ case. We will consider diagonal discriminant analysis (DLDA) (Dudoit et al., 2002), where off-diagonal elements of the covariance matrix are set to 0 and the inverse is found by inverting the variances. Shrunken linear discriminant analysis (SLDA), which utilizes the inverse of an optimally shrunken covariance matrix estimate, constructed from a convex linear combination of the sample covariance matrix and the identity matrix (Ledoit and Wolf, 2004). Maximum Uncertainty Linear Discriminant Analysis (MLDA) (Thomaz et al., 2006) estimates the covariance matrix through the maximum entropy covariance selection approach. These methods have been shown to improve LDA classification performance in the high dimensional setting by better conditioning the covariance matrix inverse (Sharma and Paliwal, 2015). Therefore, we hypothesize that they are likely to perform better than LDA with pseudo-inverse in terms of error rate.

6.4 Result

From randomizing features order 50 times and averaging the error rate, we can see that the data exhibits a double descent curve. Error peaks at random guessing 0.5 when $\gamma = 1$, and decreases monotonically in the over-parameterized regime when features are increased to 1000. The error reaches a lower rate than the under-parameterized regime at around $\gamma = 1.5$, which means high

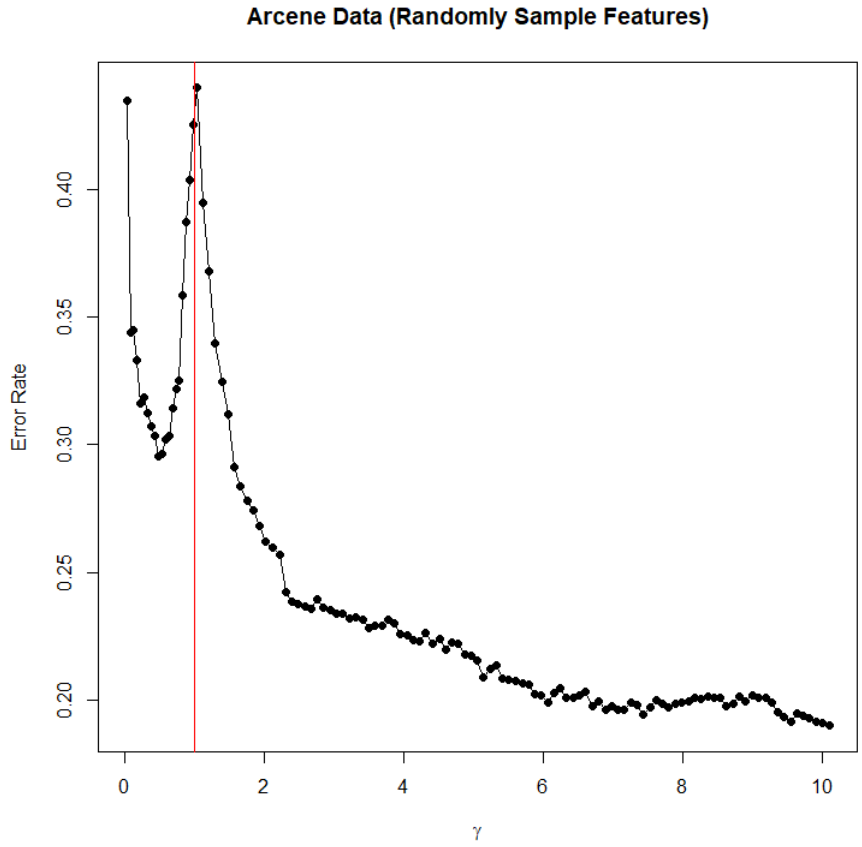


Figure 6.2: ARCENE data observed error rate for $\gamma \in (0, 10)$. Features are randomized and the result is averaged over 50 trials

over-parameterization is not necessary. There does not appear to be a second ascend in error despite the violation of normality, noise, and redundant features. The standard deviation of the error estimate of each trial is also reasonable.

6.4.1 Comparison to other methods

PCA

Through conducting PCA, the minimum error rate is reached when around 60 principal components are used. Compared to the under-parameterized regime of pseudo-LDA, the error rate is lower overall (minimum at 0.1 compared to 0.27). However, the error is similar to the over-parameterized regime of PLDA. This shows that large over-parmeterization and the effect of double descent can help PLDA achieve a similar error rate as applying dimension reduction.

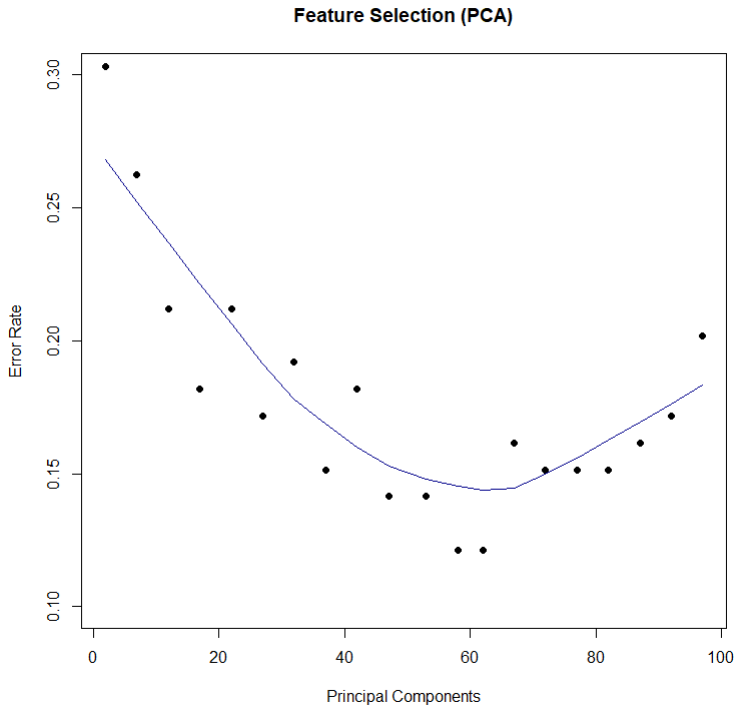


Figure 6.3: LDA Error Rate on ARCENE data with features as principal components

	Under-parameterized Minimum Error	Over-parameterized Minimum Error
PLDA	0.293 ($\gamma = 0.7$)	0.202 ($\gamma = 2.4$)
DLDA	0.303 ($\gamma = 0.7$)	0.273 ($\gamma = 3.5$)
SLDA	0.293 ($\gamma = 0.4$)	0.253 ($\gamma = 2.1$)
MLDA	0.212 ($\gamma = 0.7$)	0.253 ($\gamma = 1.1$)

Table 6.1: Error rate comparison for different High Dimensional LDA methods

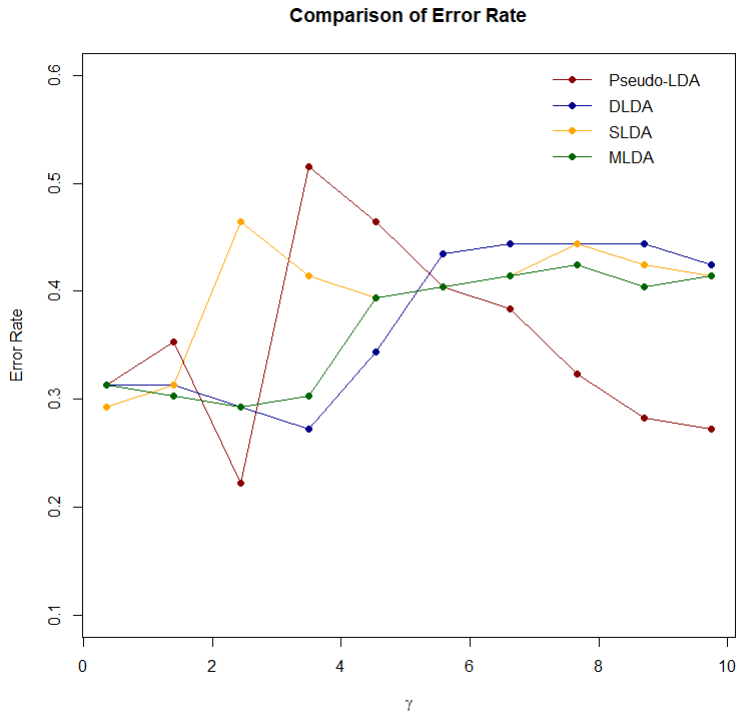


Figure 6.4: Comparison with other methods of high dimensional LDA using originally ordered features $\gamma \in (0, 10)$

Other methods

Using originally ordered features and comparing PLDA error rates to other high dimensional LDA techniques, we can see that the behaviour of the error rate is drastically different. In the underparameterized regime, PLDA has the highest error rate relative to methods like MLDA, due to ill-conditioning of covariance matrices as p approaches n. However, at large γ values, DLDA, SLDA and MLDA all converge to a high error rate ($\simeq 0.45$), whilst PLDA has a continuously decreasing error rate after a certain point. This is confirmed by the table of minimum error, pseudo-inverse's performance is most competitive in the over-parameterized regime. The overall lowest error is also achieved by PLDA.

6.5 Discussion

Overall, we have observed the double descent phenomenon on the ARCENE data, as suggested by theory and previous literature. LDA with pseudo-inverse has good performance in the high dimensional setting due to the implicit regularization of the covariance matrix despite heavy model assumption violation. The global minimum error occurs at $\gamma > 1$ and is much lower than the local minimum at $\gamma < 1$. Therefore, in practice, it is useful to consider fitting the PLDA model by increasing features as much as possible rather than selecting a small number of features to find the first minimum point. However, one should consider the nature of the dataset before applying pseudo-LDA. If the dimension is only slightly more than the sample size, it might be useful to consider other methods such as SLDA, MLDA. Feature selection such as PCA can also help LDA achieve a low error rate at a small computational cost. However, interpretability is lost as features are transformed into principal components. If γ can be scaled to more than 5, PLDA is more competitive, but the computation complexity of inverting and storing large matrices might make the method infeasible. In that case, reducing sample sizes could be considered. It is also useful to consider "ensemble methods" such as constructing different PLDA models with randomly selected features at a large γ value and averaging the prediction result. Due to the small sample size problem, it is hard to estimate the population parameter and apply theoretical prediction of error rate. However, the shape of the risk curve suggested by theory is still useful to inform the behaviour of error rate.

7 Conclusion and Future Work

7.1 Summary of Contributions

7.1.1 Theoretical Contribution

In our theoretical analysis, we derived the convergence of LDA's error rate under double asymptotic setting in the under-parameterized regime. To the best of our knowledge, we were the first to apply random matrix theory results without the gaussian assumption to this problem setting. By doing so, we uncovered the relationship between LDA's error rate and model flexibility in the double asymptotic setting.

7.1.2 Simulation Contribution

Through simulation, we were able to comprehensively demonstrate the existence of double descent and benign overfitting for pseudo-inverse LDA. In the under-parameterized regime, the error exhibits U-shaped curve and peaks around $p \simeq n$, due to ill conditioning of the covariance matrix. In the over-parameterized regime, the error either continues to decrease for indefinitely large values of γ or exhibit a local minimum before rising upwards again. This phenomenon is observed for various kinds of settings. We firstly notice that more sample sizes hurt can hurt model performance if n is close to p. We also notice that as the extent of correlation between features increase, a larger over-parameterization ratio is required to achieve global minimum error rate. Non-normality, noisy observations and redundant observations can all compromise benign overfitting by raising error rate at each value of γ.

7.1.3 Real Data Analysis

Using the ARCENE cancer classification data, we confirm that double descent can occur on real life data-set. By comparing the Pseudo-LDA method with other high dimensional classification techniques, we discovered that Pseudo-LDA has competitive performance in the $p > n$ region, which means that it is a viable method if the data set has many more features than sample size.

7.1.4 Practical considerations and avoiding Double Descent

As confirmed by LDA and other models, the double descent and benign overfitting phenomenon informs machine learning practitioners that we should rethink the way we train models. Dimension reduction is not always necessary to achieve the minimum error point as the global minimum occur with over-parameterization. This knowledge is particularly beneficial when dealing with data of limited sample size, such as medical records, gene expression etc. In practice, one should still should try various methods to determine if dimension reduction is necessary, especially to avoid multicollinearity and redundant features. Another pit-fall of the pseudo-LDA method is the high cost of computation when computing the pseudo-inverse in ultra high dimensions.

Furthermore, it is worth noting that observing double descent is not necessarily desirable as one still has to avoid the peaking phenomenon. It has been observed that other classification methods such as 1-st nearest neighbour or nearest mean can have monotonically decreasing error as $\gamma \to \infty$ Duin (1995). Due to time constraints, this paper does not focus on methods of eliminating double descent or improving error rate. Previous literature have shown success in this regard. Hoyle (2010) noted that the divergence of error when pseudo-inverse is used as $n \to p$ can be avoided by simple regularization or shrinking the population covariance matrix towards a simpler model form with well-behaved estimates. Hoyle also found that the use of bagging (estimating covariance matrix using bootstrap samples) and the Random Subspace Method (selecting a random subset of original features) can ameliorate the peaking phenomenon. Bias correction for LDA and RLDA discriminant rule has also been proposed by

Wang and Jiang (2018).

7.2 Suggestions for Future Work

Due to time constraint, we were not able to develop the theoretical result for the case where $\gamma \in (1, \infty)$, This region is of interest to the double descent as it would allow us to better compare the two minima error location. Simulation results using the theoretical results and scaling p to larger values should also be produced to validate the result. Furthermore, we note that the theoretical error convergence of LDA is still within the normal cumulative distribution, despite not having the normal assumption in our derivation. Although this is not a uncommon restraint in statistics, but we can attempt to prove that the limit of the difference between the error with normal cumulative distribution and an arbitrary cumulative distribution converges to 0. This would make the theoretical result more standalone and complete. Furthermore, we could generalize the result to include the case where sample size for each groups are not equal.

We can further extend the theoretical analysis to include cases where group sample sizes are not equal, or to variants of LDA such as quadratic discriminant analysis, mixture discriminant analysis etc. We may also consider the multi-class classification problem. Furthermore, we can consider different data structures such as non-i.i.d. time series data.

A Appendix

A.1 Appendix A

Applying theorem 3.6 of Bai and silverstein (2009) and the Marcenko-Pastur law:

Using standard M-P law, $\sigma^2 = I$, let $x = 1 + \gamma + 2\sqrt{\gamma}cos(w)$, $\zeta = e^{iw}$, where $i = \sqrt{-1}$.

$$\int_{(1-\sqrt{\gamma})^2}^{(1+\sqrt{\gamma})^2} \frac{1}{x^2} dF_\gamma(x) = \int_{(1-\sqrt{\gamma})^2}^{(1+\sqrt{\gamma})^2} \frac{1}{2\pi x^3 \gamma} \sqrt{((1+\sqrt{\gamma})^2 - x)(x - (1-\sqrt{\gamma})^2)} \tag{A.1}$$

$$= \int_0^\pi \frac{2}{\pi} \frac{sin^2(w)}{(1+\gamma+2\sqrt{\gamma}cos(w))^3} dw \tag{A.2}$$

$$= \frac{1}{\pi} \int_0^{2\pi} \frac{((e^{iw} - e^{-iw})/2i)^2}{(1+\gamma+\sqrt{\gamma}(e^{iw}+e^{-iw}))^3} dw \tag{A.3}$$

$$= -\frac{1}{4\pi i} \oint_{|\zeta|=1} \frac{(\zeta - \zeta^{-1})^2}{\zeta(1+\gamma+\sqrt{\gamma}(\zeta+\zeta^{-1}))^3} d\zeta \tag{A.4}$$

$$= -\frac{1}{4\pi i} \oint_{|\zeta|=1} \frac{(\zeta^{-1}(\zeta^2-1))^2}{\zeta(1+\gamma+\sqrt{\gamma}\zeta^{-1}(\zeta^2+1))^3} d\zeta \tag{A.5}$$

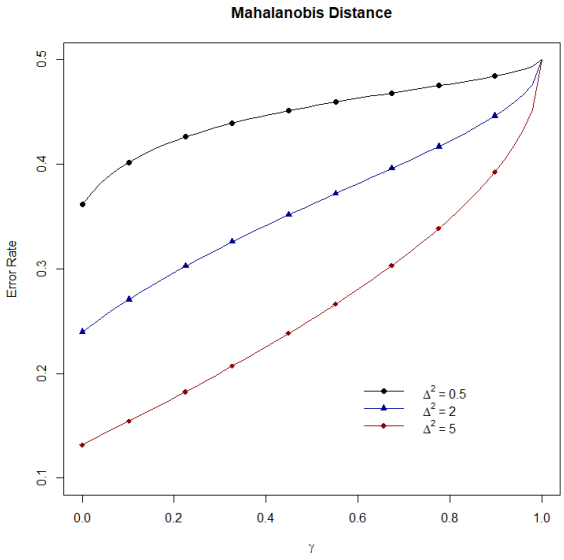
$$= -\frac{1}{4\pi i} \oint_{|\zeta|=1} \frac{(\zeta^2-1)^2}{(\zeta+\zeta\gamma+\sqrt{\gamma}(\zeta^2+1))^3} d\zeta \tag{A.6}$$

$$= -\frac{1}{4\pi i} \oint_{|\zeta|=1} \frac{(\zeta^2-1)^2}{(\sqrt{\gamma}(\zeta+\frac{1}{\gamma})(\zeta+\sqrt{\gamma}))^3} d\zeta \tag{A.7}$$

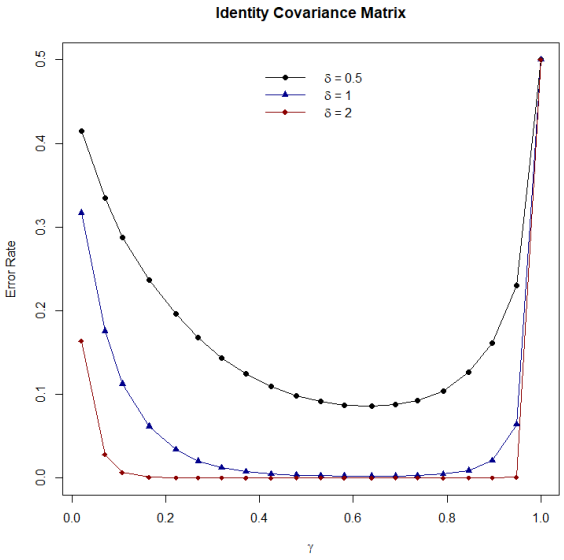
The expression evaluates to $\zeta = -\frac{1}{\sqrt{\gamma}}$ and $\zeta = -\sqrt{\gamma}$. Since $\gamma < 1$, only $-\sqrt{\gamma}$ is within the region of integration. Apply Cauchy's residue theorem for the pole of order $m = 3$, we get the desired result.

A.2 Appendix B

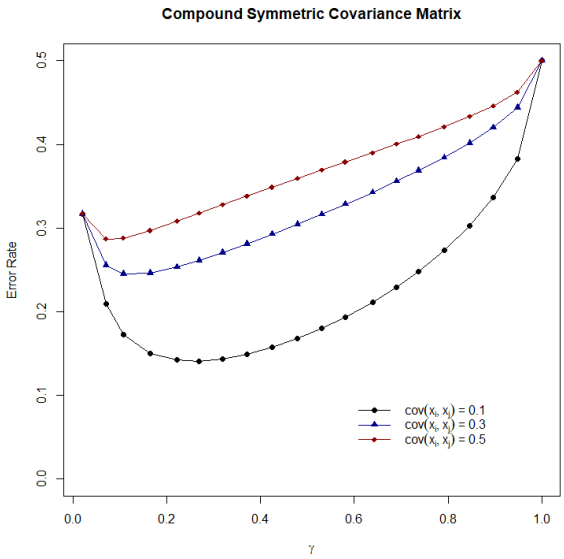
Error Rate under normal assumption for the Under-parameterized : Theorem 1 of Wang and Jiang (2018).



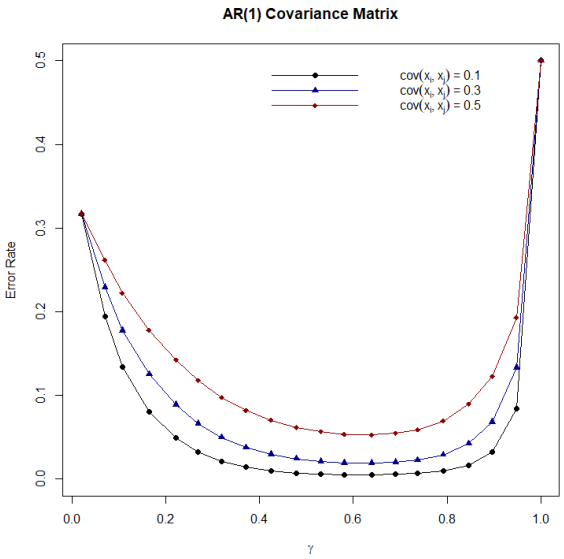
(a) Relationship of γ and Error for varying levels of squared Mahalanobis Distance Δ^2



(b) Identity covariance matrix with varying population group mean differences. $\delta = \mu_1 - \mu_2, \Sigma = I$



(c) Compound symmetric covariance matrix with different covariances



(d) AR(1) Covariance matrices with different covariances

Figure A.1: Theoretical model error in under-parameterized regime

Bibliography

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

Bai, Z., H. Liu, and W.-K. Wong (2009). Enhancement of the applicability of markowitz's portfolio optimization by utilizing random matrix theory. *Mathematical Finance 19*(4), 639–667.

Bai, Z. and J. W. Silverstein (2010). *Spectral analysis of large dimensional random matrices*, Volume 20. Springer.

Bai, Z. D., B. Q. Miao, and G. M. Pan (2007). On asymptotics of eigenvectors of large sample covariance matrix. *The Annals of Probability 35*(4), 1532–1572.

Bai, Z. D. and Y. Q. Yin (1993). Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix. *The Annals of Probability 21*(3), 1275 – 1294.

Bartlett, P. L., P. M. Long, G. Lugosi, and A. Tsigler (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences 117*(48), 30063–30070.

Belkin, M., D. Hsu, S. Ma, and S. Mandal (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences 116*(32), 15849–15854.

Bian, W. and D. Tao (2014). Asymptotic generalization bound of fisher's linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 36*(12), 2325–2337.

Bickel, P. J. and E. Levina (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli 10*(6), 989 – 1010.

Buschjäger, S. and K. Morik (2021). There is no double-descent in random forests.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association 106*(496), 1566–1577.

Chatterji, N. S. and P. M. Long (2021). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research 22*, 1–30.

Chen, D., X. Cao, F. Wen, and J. Sun (2013). Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3025–3032.

Cheng, J., M. Chen, H. Liu, T. Zhao, and W. Liao (2022). High dimensional binary classification under label shift: Phase transition and regularization. Technical report, arXiv preprint.

Cheng, Y. (2004). Asymptotic probabilities of misclassification of two discriminant functions in cases of high dimensional data. *Statistics & Probability Letters 67*(1), 9–17.

Deev, A. D. (1970). Representation of statistics of discriminant analysis, and asymptotic expansion when space dimensions are comparable with sample size. *Dokl. Akad. Nauk SSSR 195*(457), 759–762.

Deng, Z., A. Kammoun, and C. Thrampoulidis (2021). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA 11*(2), 435–495.

Dobriban, E. and S. Wager (2018). High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics 46*(1), 247–279.

Dudoit, S., J. Fridlyand, and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association 97*(457), 77–87.

Duin, R. (2000). Classifiers in almost empty spaces. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, Volume 2, pp. 1–7.

Duin, R. P. (1995). Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, pp. 957 – 964. Uppsala, Sweden: Springer.

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B 74*(4), 745–771.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association 84*(405), 165–175.

Fujikoshi, Y. (2000). Error bounds for asymptotic approximations of the linear discriminant function when the sample sizes and dimensionality are large. *Journal of Multivariate Analysis 73*(1), 1–17.

Fukunaga, K. (2013). *Introduction to Statistical Pattern Recognition*. Computer science and scientific computing. Elsevier Science.

Guo, Y., T. Hastie, and R. Tibshirani (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics 8*(1), 86–100.

Guyon, I., S. Gunn, A. Ben-Hur, and G. Dror (2008). Arcene. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C58P55.

Hamsici, O. C. and A. M. Martinez (2008). Bayes optimality in linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*(4), 647–657.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics 50*(2), 949–986.

Hastie, T. and R. Tibshirani (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society: Series B 58*(1), 155–176.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. New York: Springer.

Hills, M. (1966). Allocation rules and their error rates. *Journal of the Royal Statistical Society. Series B (Methodological) 28*(1), 1–31.

Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics 2*(3), 360 – 378.

Hoyle, D. C. (2010). Accuracy of pseudo-inverse covariance learning—a random matrix theory analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 33*(7), 1470–1481.

Johnstone, I. M. and D. M. Titterington (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 367*(1906), 4237–4253.

Karoui, N. E. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: Risk underestimation. *The Annals of Statistics 38*(6), 3487–3566.

Ledoit, O. and S. Péché (2011, October). Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields 151*(1), 233–264.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Li, Z. and J. Yao (2016, April). On two simple and effective procedures for high dimensional classification of general populations. *Statistical Papers 57*(2), 381–405.

Marčenko, V. A. and L. A. Pastur (1967). DISTRIBUTION OF EIGENVALUES FOR SOME SETS OF RANDOM MATRICES. *Mathematics of the USSR-Sbornik 1*(4), 457–483.

MARDIA, K. V. (1970, December). Measures of multivariate skewness and kurtosis with applications. *Biometrika 57*(3), 519–530. _eprint: https://academic.oup.com/biomet/article-pdf/57/3/519/702615/57-3-519.pdf.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance 7*(1), 77–91.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

Michie, D., D. Spiegelhalter, and C. Taylor (1994). *Machine Learning, Neural and Statistical Classification*. New York: Ellis Horwood.

Montanari, A., F. Ruan, Y. Sohn, and J. Yan (2019). The generalization error of max-margin linear classifiers: Benign overfitting and high-dimensional asymptotics in the overparametrized regime. Technical report, arXiv.

Moore, E. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of American Mathematical Society 26*, 394–395.

Nakkiran, P., G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever (2021). Deep double descent: where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment 2021*(12), 124003.

Nakkiran, P., P. Venkat, S. Kakade, and T. Ma (2021). Optimal regularization can mitigate double descent. Technical report, arxiv.

Okamoto, M. (1963). An Asymptotic Expansion for the Distribution of the Linear Discriminant Function. *The Annals of Mathematical Statistics 34*(4), 1286 – 1301.

Pan, G. M. and W. Zhou (2011). Central limit theorem for Hotelling's T2 statistic under large dimension. *The Annals of Applied Probability 21*(5), 1860 – 1910.

Paul, D. and A. Aue (2014). Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference 150*, 1–29.

Penrose, R. (1955). A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society 51*(3), 406–413.

Pillo, P. J. D. (1976). The application of bias to discriminant analysis. *Communications in Statistics - Theory and Methods 5*(9), 843–854.

Qiao, Z., L. Zhou, and J. Z. Huang (2008). Effective linear discriminant analysis for high dimensional, low sample size data. In *Proceeding of the world congress on engineering*, Volume 2, pp. 2–4. Citeseer.

Raudys, S. (1967, 01). On determining the training sample size of a linear classifier. *Computing Systems 28*, 79–87.

Raudys, S. and R. P. W. Duin (1998). Expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters 19*(5), 385–392.

Saranadasa, H. (1993). Asymptotic expansion of the misclassification probabilities of d- and a-criteria for discrimination from two high dimensional populations using the theory of large dimensional random matrices. *Journal of Multivariate Analysis 46*(1), 154–174.

Shao, J., Y. Wang, X. Deng, and S. Wang (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics 39*(2), 1241–1265.

Sharma, A. and K. K. Paliwal (2015). Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics 6*, 443–454.

Sifaou, H., A. Kammoun, and M.-S. Alouini (2020). High-dimensional linear discriminant analysis classifier for spiked covariance model. *Journal of Machine Learning Research 21*, 1–24.

Silverstein, J. (1995). Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis 55*(2), 331–339.

Thomaz, C. and D. Gillies (2005). A maximum uncertainty lda-based approach for limited sample size problems - with application to face recognition. In *XVIII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'05)*, pp. 89–96.

Thomaz, C. E., E. C. Kitani, and D. F. Gillies (2006). A maximum uncertainty lda-based approach for limited sample size problems–with application to face recognition. *Journal of the Brazilian Computer Society 12*, 7–18.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science 18*(1), 104–117.

Tracy, C. A. and H. Widom (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics 159*(1), 151–174.

Wang, C. and B. Jiang (2018). On the dimension effect of regularized linear discriminant analysis. *Electronic Journal of Statistics 12*(2), 2709–2742.

Wigner, E. P. (1951). On the statistical distribution of the widths and spacings of nuclear resonance levels. *Mathematical Proceedings of the Cambridge Philosophical Society 47*(4), 790–798.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics 62*(3), 548–564.

Wyman, F. J., D. M. Young, and D. W. Turner (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition 23*(7), 775–783.

Xing, E. P., M. I. Jordan, and R. M. Karp (2001). Feature selection for high-dimensional genomic microarray data. In *ICML'01: Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 601–608.

Yin, Y., Z. Bai, and P. Krishnaiah (1988). On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. *Probability Theory and Related Fields 78*(4), 509–521.

Yu, H. and J. Yang (2001). A direct lda algorithm for high-dimensional data — with application to face recognition. *Pattern Recognition 34*(10), 2067–2070.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals (2016). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM 64*(3), 107–115.

Zollanvari, A., U. M. Braga-Neto, and E. R. Dougherty (2011). Analytic study of performance of error estimators for linear discriminant analysis. *IEEE Transactions on Signal Processing 59*(9), 4238–4255.

Šarūnas Raudys and D. M. Young (2004). Results in statistical discriminant analysis: a review of the former soviet union literature. *Journal of Multivariate Analysis 89*(1), 1–35.