HIGH DIMENSIONAL CLASSIFICATION FOR SPATIALLY DEPENDENT DATA WITH APPLICATION TO NEUROIMAGING

By

Yingjie Li

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2018

ABSTRACT

HIGH DIMENSIONAL CLASSIFICATION FOR SPATIALLY DEPENDENT DATA WITH APPLICATION TO NEUROIMAGING

By

Yingjie Li

The methods developed in this thesis are motivated by how to use structure Magnetic resonance imaging (MRI) data to predict Alzheimer's disease (AD) or to discriminate between healthy subjects and AD patients. Imaging data is a typical example of spatially dependent data where the correlation between data points collected at various voxels (pixels) can be described by proximity. Also, it is high dimensional data since the number of voxels is extremely high comparing to the number of subjects.

The first piece of work considers use longitudinal volumetric MRI data of five regions of interest (ROIs), which are known to be vulnerable to Alzheimer's disease (AD) for prediction. A longitudinal data prediction method based on functional data analysis is applied for identifying when an early prediction can reasonably be made and determining whether one ROI is superior with regard to predicting progression to AD over others. By adopting statistically validated procedures, we compared the prediction performance based on individual ROIs as well as their combinations. For all the models, the results show that the overall one year, two years, three years in advance prediction accuracy is above 80%. MCI converter subjects can be correctly detected as early as two years prior to conversion.

The second piece of work considers use voxel level MRI data for classification of AD patients and healthy subjects. A supervised learning method based on the linear discriminant analysis (LDA) was developed for high dimensional spatially dependent data. The theory shows that the method proposed can achieve consistent parameter estimation, consistent

features selection, and asymptotically optimal misclassification rate. The extensive simulation study showed a significant improvement in classification performance under spatial dependence. We applied the proposed method to voxel level MRI data for classification. The classification performance is superior compared to other comparable methods.

I dedicate this dissertation to my family, especially to my dearest grandmother Chen, Dezhen.

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to my advisor Professor Tapabrata Maiti, for his invaluable assistance, constructive guidance and immense knowledge. His vision on statistics and his enthusiasm on research encouraged me all the time. I am fully indebted to him for his understanding, patient and encouragement. It is impossible to have this thesis completed and comprehensive without his continuous support to me and to my family.

I would like to thank Professor David Zhu, for sharing his precious experience of working on brain imaging data, introducing many of the interesting practical problems in Alzheimer's disease, and also for serving as one of my committee members. I also wish to thank Professor Chae Young Lim and Professor Pingshou Zhong for serving on my dissertation committee. I am extremely grateful for their constructive advice in this thesis work.

I would also thank the entire faculty and staff members in the Department of Statistics and Probability who have helped me during my study in Michigan State University.

Last I would like to thank my parents and my sister for loving me and supporting me at every stage of my life. Many thanks go to my husband Yuzhen for standing by my side always. I give my special thanks to my daughter Aurora, who brought a wonderful light to my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 Motivation

Brain imaging provides a non-invasive way to observe the human central nervous system. Many researches are working on improving imaging technologies, data analysis and the application of imaging to investigate neurological disorders (including Alzheimer's disease (AD), Parkinson's disease and dementia). AD is the most common form of dementia. Many researches are focusing on find better ways to treat the disease, delay its onset, and prevent it from developing. Also, it is of great interest in developing objective biologically based methods to diagnose and predict Alzheimer's disease, track the progression, and monitor the efficacy of treatment. Brain atrophy is a primary pathologic process in AD due to widespread neuronal death (see, e.g., [28]). As a non-invasive, widely available and cost-effective way for obtaining brain imaging, MRI data plays an important role as a potential biomarker to monitor atrophy progression and thereby the progression of AD. Figure 1.1 gives an example of MRI of subject with AD and that of normal subject (NL).

This work is motivated by how to use structure Magnetic resonance imaging (sMRI) data of the brain to discriminate between the healthy subjects and the AD patients and to predict Alzheimer's disease (AD) progress. Imaging data is a typical example of spatial dependent data where the relationship between the data points collected at various voxels (pixels) can be described by proximity. Also, it is high dimensional data since the number of voxels

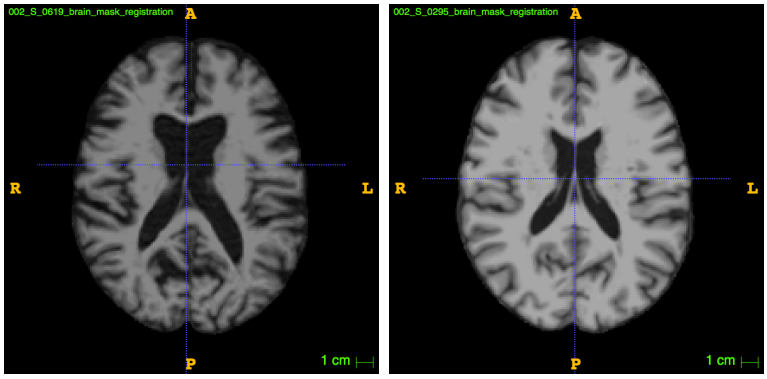


Figure 1.1 left: T1-weighted MRI of AD subject; right: TI-weighted MRI of NL subject.

is extremely high comparing to the number of subjects. If consider the signal from each voxel as a feature, we might have millions of features for classification. Motivated by the characteristics of the brain imaging data, the goal of this thesis is to develop classification methods for high dimensional data with spatial dependency.

1.2 Spatial statistical models

Spatial statistical models are widely applied to many scientific disciplines, including geology, agriculture, climatology, ecology, epidemiology, forestry, astronomy, atmospheric science, real estate, brain imaging, and any discipline that the spatial dependence among observations cannot be ignored in data analysis.

According to [42], there are three types of spatial data.

1. Geostatistical data. Let D be a continuous domain in the d-dimensional Euclidean space R^d. Let $\boldsymbol{Y}$ be the variable concerned in the study, such as the air temperature at different locations in US. D is a continuous domain means $\boldsymbol{Y}$ can be observed everywhere within D, that is between any two locations s_i and s_j, theoretically there are infinite locations where $\boldsymbol{Y}$ are well defined. We use spatial process $\{\boldsymbol{Y}(\mathbf{s}), \mathbf{s} \in D\}$

to model the variable $\boldsymbol{Y}$. If data are collected at some locations $s_1, ..., s_n$ in D, say $Y(s_1), ..., Y(s_2)$, the dataset is called geostatistical data.

2. Lattice data. Lattice data are spatial data where the domain D is fixed and discrete, in other words, non-random and countable. For example, lattice data could be data collected by ZIP code, prime rate of a county, or remotely sensed data reported by pixels. The distinction between geostatistical data and lattice data are not always clearcut. For example, the imaging data with very low resolution can be considered as lattice data; the imaging data with very high resolution can be considered as geostatistical data; those with moderate resolution could be considered as either one, determined by the goal of the study. For the convenience in the notation, we use $\{\boldsymbol{Y}(s_i), s_i \in D, i = 1, 2, ...\}$ to represent lattice data.

3. Point patterns. Geostatistical and lattice data have in common the fixed, non-stochastic domain. An important feature of point pattern data is the random domain. Locations of avian flu over the world, locations of long leaf pines in a natural forest are examples of point pattern.

Imaging data can be considered as geostatistical data. Geostatistical data could be modeled by spatial random field $\{y(s), s \in D \subset R^d\}$ such that

$$y(s) = \mu(s) + \epsilon(s)$$

where $\mu(s)$ is the mean effect function and $\epsilon(s)$ is the random noise. Assume the error term $\epsilon(s) : s \in D$ is a Gaussian random field with mean zero and a covariance function $\gamma(s, s'; \boldsymbol{\theta}) = Cov(\epsilon(s), \epsilon(s'); \boldsymbol{\theta})$, where s, $s' \in D$ and $\boldsymbol{\theta}$ is a $q \times 1$ vector of covariance

function parameter which could be estimated from the data. If we further assume the random field $\epsilon(s)$ is isotropic and stationary, the covariance function could be represented as $\gamma(h;\boldsymbol{\theta}) = cov(\epsilon(s), \epsilon(s');\boldsymbol{\theta})$, where $h = \left\|s - s'\right\|$ is the Euclidean distance between s and s'.

There are many ways to model the covariance function $\gamma(h;\boldsymbol{\theta})$. A widely used family of covariance function is the *Matérn* covariance function. It is defined as:

$$\gamma(h;\sigma^2, c, \nu, r) := \sigma^2(1-c)\frac{2^{1-\nu}}{\Gamma(\nu)}(h/r)^{\nu}K_{\nu}(h/r) \tag{1.2.1}$$

where $K_{\nu}(\cdot)$ is a modified Bessel function of the second kind and $\sigma^2 > 0$ is the variance, $0 \leq c \leq 1$ is a nugget effect, $\nu > 0$ is the scale and smoothness parameter (see, e.g., [17]). First, The *Matérn* covariance function is isotropic. The correlation decreases when the distance h increases. Second, when ν increases, the smoothness of the random field increases, the *Matérn* covariance function converges to Gaussian covariance function $\gamma(h;\sigma^2, c, r) = \sigma^2(1-c)exp(-h^2/r^2)$ as $\nu \to \infty$. Last, if $\nu = \frac{1}{2}$, (1.2.1) is reduced to the well known exponential covariance function $\gamma(h;\sigma^2, c, r) = \sigma^2(1-c)exp(-h/r)$. r is called the range parameter since it measures the distance at which the correlation have decreased below certain threshold.

Consider the spatial statistical model with replications. Assume there are n independent samples of the spatial random field: $y_1(s), y_2(s), ..., y_n(s)$. Suppose any sample of the spatial random field, there are observations at p discrete sites $s_1, ..., s_p \in D$. Let $Y_{ij} = y_i(s_j)$ be the observation at the jth site from the ith sample of the spatial random field. For $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$, the model can be represented by:

$$Y_{ij} = \mu_j + \epsilon_{ij} \tag{1.2.2}$$

where $\mu_j = \mu(s_j)$ is the mean effect at the jth location and $\epsilon_{ij} = \epsilon_i(s_j)$ is the corresponding Gaunssian random noise at the jth location for the ith sample. We can write the model in matrix notation. For $i = 1, 2, ..., n$, let $\boldsymbol{Y}_i = (Y_{i1}^T, ..., Y_{ip}^T)^T$ be a p-dimensional vector. Then $\boldsymbol{Y}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i$, where $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ is the mean vector and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_p)^T$ has multivariate normal distrubition $N(0, \Sigma)$. Since $\epsilon(s)$ has a covariance function $\gamma(h; \boldsymbol{\theta})$, the covariance matrix Σ can be represented by $\Sigma(\boldsymbol{\theta}) = [\gamma(h_{ij}; \boldsymbol{\theta})]_{i,j=1}^p$, where $h_{ij} = \left\| s_i - s_j \right\|$, i.e. the $(i, j)th$ entry of $\Sigma(\boldsymbol{\theta})$ is $\gamma(h_{ij}; \boldsymbol{\theta})$.

Maximum likelihood estimation (MLE) are often used to estimate parametric model (see, e.g., [37]). Considering the model in (1.2.2), $\boldsymbol{Y}_i \sim N(\mu, \Sigma(\boldsymbol{\theta}))$. Let $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, ..., \boldsymbol{Y}_n^T)$, the MLE of $(\boldsymbol{\mu}, \boldsymbol{\theta})$ is obtained by maximizing the log likelihood function,

$$\log\, l(\boldsymbol{\mu}, \boldsymbol{\theta}; \boldsymbol{Y}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma(\boldsymbol{\theta})| - \tfrac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{Y}_i - \boldsymbol{\mu})$$

The unique $\boldsymbol{\mu}$ to maximize $\log l(\boldsymbol{\mu}, \boldsymbol{\theta}; \boldsymbol{Y})$ is: $\hat{\boldsymbol{\mu}}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Y}_i = \bar{\boldsymbol{Y}}$. Plug in $\hat{\boldsymbol{\mu}}$ to $\log\, l(\boldsymbol{\mu}, \boldsymbol{\theta}; \boldsymbol{Y})$, we obtain

$$L(\boldsymbol{\theta}; \boldsymbol{Y}) = -\frac{n}{2}\log|\Sigma(\boldsymbol{\theta})| - \tfrac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}})^T\Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}})$$

$\hat{\boldsymbol{\theta}}_{MLE}$ can be obtained by maximizing $L(\boldsymbol{\theta}; \boldsymbol{Y})$. Usually there's no closed-form solution. Numerical methods could be applied to obtain the solution.

1.3 Linear discriminant analysis (LDA) for high dimensional data

Consider the p-dimensional discriminant problem between two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. According to some classification rule $T(\mathbf{x}) : R^d \to \{1, 2\}$, a new observation $\boldsymbol{X} = \mathbf{x}$ can be classified into class $\mathcal{C}_1$ ($T(\boldsymbol{X}) = 1$) or $\mathcal{C}_2$ ($T(\boldsymbol{X}) = 2$). If $\boldsymbol{X} \in \mathcal{C}_1$, the misclassification rate is the probability that it is classified into class $\mathcal{C}_2$, i.e. $P(T(\boldsymbol{X}) = 2|\boldsymbol{X} \in \mathcal{C}_1)$. Similary, $P(T(\boldsymbol{X}) = 1|\boldsymbol{X} \in \mathcal{C}_2)$ is the misclassification rate if $\boldsymbol{X} \in \mathcal{C}_2$. Here $\boldsymbol{X}$ is a p-dimensional random variable. By abuse of notation we use the notation: $\boldsymbol{X} \in \mathcal{C}_k$ to denote that the observation $\boldsymbol{X}$ is from class $\mathcal{C}_k$.

It can be shown that the optimal classifier in terms of minimizing the misclasification rate is known as the Bayes rule, which classify the new observation into the most probable class, using the posterior probability of the observation (see Chapter 2 in [23]). Suppose $f_k(\mathbf{x})$ is the conditional density of an observation $\boldsymbol{X}$ in class $\mathcal{C}_k$, ($k = 1, 2$). Let π_k be the prior probability of class k with $\pi_1 + \pi_2 = 1$. By Bayes theorem, the posterior probability of an observation $\boldsymbol{X} = \mathbf{x}$ in each class is:

$$P(\boldsymbol{X} \in \mathcal{C}_k|\boldsymbol{X} = \mathbf{x}) = \frac{f_k(\mathbf{x})\pi_k}{f_1(\mathbf{x})\pi_1 + f_2(\mathbf{x})(1 - \pi_1)}$$

There are many ways to model the class densities. Suppose the densities for both classes are multivariate Gaussian $N(\boldsymbol{\mu}_1, \Sigma)$ and $N(\boldsymbol{\mu}_2, \Sigma)$ respectively, where $\boldsymbol{\mu}_k$ ($k = 1, 2$) are the class mean vectors and Σ is the common positive definite covariance matrix. Then the

density of an observation $\boldsymbol{X} = \mathbf{x}$ from $\mathcal{C}_k$ is:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}}|\Sigma|^{1/2}e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_k))}$$

Under this assumption, the Bayes rule assigns $\boldsymbol{X} = \mathbf{x}$ in to $\mathcal{C}_1$ if $\pi_1 f_1(\mathbf{x}) \geq \pi_2 f_2(\mathbf{x})$. Equivalently, $\mathbf{x}$ is assigned to $\mathcal{C}_1$ if

$$\log\frac{\pi_1}{\pi_2} + (\mathbf{x} - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq 0$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Note that this classifier is linear in $\mathbf{x}$.

In practice, the parameters of the Gaussian distribution are unknown. We can estimate them using training data. Suppose $\boldsymbol{Y}_{k1}, ..., \boldsymbol{Y}_{kn_k}$ are from class $\mathcal{C}_k$, where $k \in \{1, 2\}$ and $\boldsymbol{Y}_{kj} \in \mathbb{R}^p$ are independent and identically distributed as $N_p(\boldsymbol{\mu}_k, \Sigma(\boldsymbol{\theta}_0))$, where $\boldsymbol{\mu}_k = (\mu_{k1}, ..., \mu_{kp})^T$, n_k is the sample size for class $\mathcal{C}_k$. $\Sigma(\boldsymbol{\theta})$ is the covariance matrix with parameter $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Let $n = n_1 + n_2$ be the total sample size. Assume $\frac{n_1}{n} \to \pi$, $0 < \pi < 1$ as $n \to \infty$. p is depending on n. Assume the two classes have equal prior probability for the two classes, i.e. both the probability that a new observation is from class $\mathcal{C}_1$ and $\mathcal{C}_2$ are $\frac{1}{2}$.

Consider the classification rule $\hat{\delta}$:

$$\hat{\delta}(\boldsymbol{X}) = (\boldsymbol{X} - \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2})^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} > 0 \tag{1.3.1}$$

where $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$, $\hat{\Sigma}$ and $\hat{\boldsymbol{\Delta}}$ are estimates of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, Σ and $\boldsymbol{\Delta}$ from the data, where $\boldsymbol{\Delta} = (\Delta_1, ..., \Delta_p)^T = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is the difference of the two classes in mean. A new observation $\mathbf{x}$ is classified into class $\mathcal{C}_1$ if $\hat{\delta}(\mathbf{x}) > 0$ and $\mathcal{C}_2$ otherwise. Let $\Theta = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\theta})$. If the new

observation $\boldsymbol{X}$ comes from $\mathcal{C}_1$, then the conditional misclassification rate of $\hat{\delta}$ is:

$$W_1(\hat{\delta};\Theta) = P(\hat{\delta}(\boldsymbol{X}) \leq 0 | \boldsymbol{X} \in \mathcal{C}_1, \boldsymbol{Y}_{ki}, i = 1, 2, ..., n_k, k = 1, 2) = 1 - \Phi(\Psi_1) \tag{1.3.2}$$

where

$$\Psi_1 = \frac{(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}})^T \hat{\Sigma} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}{\sqrt{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}} \tag{1.3.3}$$

Similarly we can define the error rate for observations from $\mathcal{C}_2$. If a new observation $\boldsymbol{X}$ comes from class $\mathcal{C}_2$, the conditional misclassification rate of $\hat{\delta}$ is:

$$W_2(\hat{\delta}, \Theta) = \mathbb{P}(\hat{\delta}(\boldsymbol{X}) > 0 | \boldsymbol{X} \in \mathcal{C}_2, \boldsymbol{Y}_{ki}, k = 1, 2; i = 1, ..., n_k) = \Phi(\Psi_2) \tag{1.3.4}$$

where

$$\Psi_2 = \frac{(\boldsymbol{\mu}_2 - \hat{\boldsymbol{\mu}})^T \hat{\Sigma} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}{\sqrt{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T \hat{\Sigma}^{-1} \Sigma \hat{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}}, \tag{1.3.5}$$

Since we assumed the equal prior probability for the two classes, the overall misclassification rate is :

$$W(\hat{\delta};\Theta) = \tfrac{1}{2}(W_1(\hat{\delta};\Theta) + W_2(\hat{\delta};\Theta)) \tag{1.3.6}$$

If $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and Σ are known, the optimal classification rule is Bayes rule, which classifies

a new observation $\mathbf{X} = \mathbf{x}$ into class $\mathcal{C}_1$ if

$$\delta(\mathbf{x}) = (\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})^T \Sigma^{-1} \boldsymbol{\Delta} > 0, \tag{1.3.7}$$

Bayes rule has the smallest misclassification rate. If there's a new observation $\boldsymbol{X}$ from class $\mathcal{C}_1$, since $\boldsymbol{X}$ has normal distribution $N(\boldsymbol{\mu}_1, \Sigma(\boldsymbol{\theta}))$, we can calculate that the conditional misclassification rate of Bayes rule δ is

$$W_1(\delta; \Theta) = W_2(\delta; \Theta) = 1 - \Phi(\frac{\sqrt{C_p}}{2}) \tag{1.3.8}$$

where $C_p = \boldsymbol{\Delta}^T \Sigma^{-1}(\boldsymbol{\theta}) \boldsymbol{\Delta}$ and $\Phi(\cdot)$ is the standard Gaussian distribution function.

The overall misclassification rate of Bayes rule is:

$$W(\delta; \Theta) = \tfrac{1}{2}(W_1(\delta; \Theta) + W_2(\delta; \Theta)) = 1 - \Phi(\frac{\sqrt{C_p}}{2})$$

Since Bayes rule has the smallest misclassification rate, we write $W_{OPT} = 1 - \Phi(\frac{\sqrt{C_p}}{2})$ as the optimal misclassification rate.

In practice the parameters $\Sigma(\boldsymbol{\theta})$, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ are unknown. We need to estimate them using data. The linear discriminant analysis (LDA) estimates the parameters with sample estimates ($k = 1, 2$):

$$\hat{\boldsymbol{\mu}}_k = \sum_{i=1}^{n_k} \boldsymbol{Y}_{ki} / n_k = \bar{\boldsymbol{Y}}_{k\cdot} \tag{1.3.9}$$

$$\hat{\Sigma} = \sum_k \sum_i (\boldsymbol{Y}_{ki} - \hat{\boldsymbol{\mu}}_k)^T (\boldsymbol{Y}_{ki} - \hat{\boldsymbol{\mu}}_k) / (n_1 + n_2 - 2) \tag{1.3.10}$$

Then LDA classifies $\boldsymbol{X}$ into class $\mathcal{C}_1$ if

$$\hat{\delta}_{LDA}(\boldsymbol{X}) = (\boldsymbol{X} - \frac{1}{2}(\bar{\boldsymbol{Y}}_{1\cdot} + \bar{\boldsymbol{Y}}_{2\cdot}))^T \hat{\Sigma}^{-1}(\bar{\boldsymbol{Y}}_{1\cdot} - \bar{\boldsymbol{Y}}_{2\cdot}) > 0 \tag{1.3.11}$$

However LDA is not applicable under high dimensional setting ($p \gg n$) for two reasons. First the sample covariance matrix $\hat{\Sigma}$ is singular and it is difficult to estimate the precision matrix $\Omega = \Sigma^{-1}$. Second, even though the true covariance is known, the classification performance when $p \gg n$ is poor due to the error accumulated when estimate the p-dimensional features. Fan and Fan (2008) [19], and Shao *et al.* (2011) [46] show theoretically that the variable selection is critical respectively.

Though there are many literature working on high dimensional classification to seek a better classification performance for general correlation structure in Σ (see, e.g., [5, 9, 15, 19, 20, 33, 35, 36, 43, 46, 48, 51, 52, 55]), but still not much has been talked about the high dimensional classification for spatially dependent data. Here in this thesis we fill in the gap. As defined in [46], a classifer $\hat{\delta}$ is (1) asymptotically optimal if $W(\hat{\delta}; \boldsymbol{\theta})/W_{OPT} \xrightarrow{P} 1$; (2) asymptotically sub-optimal if $W(\hat{\delta}; \boldsymbol{\theta}) - W_{OPT} \xrightarrow{P} 0$, where $\xrightarrow{P}$ means convergence in probability. Our goal in this thesis is to develop a asymptotically optimal classifier.

1.4 Overview

The rest of this dissertation is as follows. As a pilot study, in Chapter 2 we consider use longitudinal volumetric MRI data of five regions of interest (ROIs), which are known to be vulnerable to Alzheimer's disease (AD) for prediction. A longitudinal data classification method based on functional data analysis is developed. It is applied for identifying when

an early prediction can reasonably be made and determining whether one ROI is superior with regard to predicting progression to AD over others. We first recover the longitudinal observations into trajectories using PACE (principal components analysis through conditional expectation; see [56]). PACE is a version of functional principal components (FPC) analysis, in which the FPC scores are estimated from conditional expectations. To check the appropriate time period for early prediction, using the idea from [24], we use only part of the observations from longitudinal data in this step. Then logistic regression with FPC scores as the explanatory variables was applied for classification. We compare the accuracy, sensitivity, and specificity of prediction based on individual ROIs as well as their combinations. For all the models, the results show that the overall one year, two years, three years in advance prediction accuracy is above 80%. MCI converter subjects can be correctly detected with a greater-than-chance probability as early as two years prior to conversion.

The data used in Chapter 2 is highly summarized data derived from volumetric MRI, which measures the size of each ROI. Summarized volumetric MRI data is a good way to measure the brain atrophy. However potentially there's a lot of information lost due to the ignorance of the spatial pattern. How to do classification incorporating the spatial dependency is of great interest for not only the brain imaging data, but also for the imaging data from a wide range of source.

In Chapter 3 we consider use voxel level MRI data for classification of AD patients and healthy subjects. This data is characterized with high dimension and spatial dependency. We develop a supervised learning method based on linear discriminant analysis for high dimensional spatially dependent data. To solve the two issues in standard LDA for high dimensional data, first, we incorporate the information of spatial dependency among all features using spatial statistical model. Then a non-singular covariance structure can be de-

rived by maximum likelihood estimation. Second, penalized maximum likelihood estimation (PMLE) is applied for selecting important features and estimating parameters simultaneously. To reduce computation load, the tapering technique is applied. Additionally, we develop the theories which show that the method proposed can achieve consistent parameter estimation, consistent feature selection, and asymptotically optimal misclassification rate. Extensive simulation study shows a significant improvement in classification performance under spatial dependence compared to other comparable methods. In the end we apply this method in using brain imaging data for diagnosis of Alzheimer's disease.

Chapter 2

Early prediction of Alzheimer's disease using longitudinal volumetric MRI data from ADNI

2.1 Introduction

The accurate prediction of Alzheimer's disease (AD) in individuals with Mild Cognitive Impairment (MCI) is essential for clinical management and selection of potential interventions. It also plays an important role in improving the efficiency of clinical trials of AD-modifying treatments by enriching the study population with a higher proportion of individuals who will develop the disease during the trial period. Structural MRI has been a primary research tool for the development of prognostic markers to aid in disease prediction, taking advantage of brain atrophy, a primary pathologic process in AD due to widespread neuronal death (see, e.g., [28]). Many studies suggest that structural changes in the brain can be detected in the early stages of AD (see, e.g., [2, 22, 29]). Several studies analyzed brain atrophy patterns in different regions of interests (ROIs) and different disease stages (see, e.g., [2, 22, 29, 31, 32, 38]. They found that rates of atrophy are not uniform, but vary in accordance with disease stage and by region. For example, Leung *et al.* (2010) [31] found a

higher rate of hippocampal atrophy in MCI converters (MCI-c) than in MCI nonconverters (MCI-nc). Also, Fennema-Notestine *et al.* (2009) [22] found that some regions such as the mesial temporal, exhibited a linear rate of atrophy throughout both MCI and AD. Other regions such as lateral temporal, middle gyrus showed accelerated atrophy later in the disease. These results imply that brain atrophy in certain regions can be used to differentiate MCI-c and MCI-nc. Longitudinal data captures atrophy patterns that change with time at different disease stages. Compared to single point brain imaging data, longitudinal data does a better job of describing brain atrophy and likely performs better in predicting conversion from MCI to AD.

MRI offers a non-invasive, widely available and cost-effective way for obtaining imaging biomarkers. Image analysis software such as FreeSurfer has been able to provide accurate estimation of regional brain volume, cortical thickness, and even curvature. Regional brain volume has been investigated most as a potential biomarker to monitor atrophy progression and thereby the progression of AD. In this paper we studied longitudinal volumetric data of five regions of interests (ROIs) measured with MRI, including the hippocampus (H), entorhinal cortex (EC), middle temporal cortex (MTC), fusiform gyrus (FG) and whole brain (WB), in order to develop a better potential biomarker strategy for use with structural data.

Among the papers which investigate prediction of conversion from MCI to AD using longitudinal MRI data, several stand out as particularly pertinent to our present work. Misra *et al.* (2009) [40] applied high-dimensional pattern classification to longitudinal MRI scans for prediction of conversion within an average period of 15 months. An abnormality score was calculated for classification and achieved an accuracy rate of 81.5%. Zhang *et al.* (2012) [57] performed predictions on multimodal longitudinal data (i.e., MRI, PET, etc.) using a longitudinal feature selection method. Multi-kernel SVM was used for classification.

They achieved an accuracy of 78.4%, sensitivity of 79.0%, and specificity of 78.0% for at least six months ahead of the conversion from MCI to AD. In Li *et al.* (2012) [34], 262 features were calculated from longitudinal cortical thickness MRI data. They first applied mRMR (minimum redundancy maximum relevance) for feature selection, then an SVM (support vector machine) was trained for classification. Their method can detect 81.7% (AUC = 0.875) of the MCI converters six months ahead of conversion to AD. Lee *et al.* (2016) [30] used baseline plus 1-year follow-up callosal MRI scans for predicting conversion to AD in the following 2 to 6 years. Logistic regression model with fused lasso regularization was applied. They found the accuracy of prediction was 84% in females and 61% in males. Arco *et al.* (2016) [3] performed prediction based on baseline and 1-year follow-up MRI data and clinical scores (MMSE and ADAS-Cog). Feature selection based on a two sample t-test and classification based on maximum-uncertainty linear discriminant analysis were applied. They achieved a classification accuracy of 73.95%, with a sensitivity of 72.14% and a specificity of 73.77% for six months before conversion. In Adaszewski *et al.* (2013) [1], voxel-based longitudinal structure MRI data was used. Weighted based feature selection and SVM were applied for prediction. Both sensitivity and specificity were up to 65% at 1 to 4 years before conversion. Notably they found that conversion could be detected as early as four years before conversion with a sensitivity of 62%.

As mentioned in Eskildsen *et al.* (2013) [18], one factor preventing high predictive power is the heterogeneity of the MRI data due to the variability in underlying disease stage among subjects. Without standardizing the data based on disease stage, it is impossible to establish a specific pattern of atrophy at specific disease stages. Adaszewski *et al.* (2013) [1] attempted to solve this issue by aligning the data by "time of conversion". We also aligned the data by "time of conversion" in this paper. This alignment enables us to estimate the pattern of

atrophy in accordance with the stages of disease. Moreover, it makes it feasible to compare ahead of how many years we can reasonably predict the conversion.

Another source of variability in the longitudinal MRI data is that the subjects have different numbers of observations. In the ADNI cohort, some subjects have nine years of data with up to 10 longitudinal observations, but many subjects only have 1 to 3 data points. Due to the sparsity and irregularity of the data, it is not feasible for the traditional longitudinal data analysis method to make predictions and meanwhile, compare the prediction windows. In this study, we applied a technique known as the principal component analysis through conditional expectation (PACE, see [56]) to analyze the realigned longitudinal MRI data. In PACE, the mean curve and covariance structure of the biomarker along with time are obtained based on the pooled data from all individuals. In this way, longitudinal observations from each subject could be recovered as a smooth trajectory even if only one or few observations are available. Finally, once the longitudinal observations are recovered to be smooth trajectories, they could be treated as functional data. Then a functional prediction method from Hall *et al.* (2012) [24] could be employed to determine when an early decision can reasonably be made, and identify which ROI is best for prediction, using only part of the trajectory. Moreover, we also examined whether any combination of the ROIs can provide a better prediction result.

The remainder of this chapter is organized as follows: Section 2.2 presents the methods, the PACE approach for analyzing longitudinal data and the functional prediction method. Section 2.3 introduces the data that we use in analysis. Section 2.4 presents the numerical results and conclusions. The discussion is in section 2.5. More tables about results can be found in the Appendix.

2.2 Methodology

2.2.1 Functional Principal Component Analysis with longitudinal data

We first recover the longitudinal observations into trajectories using PACE (principal components analysis through conditional expectation; [56]). PACE is a version of functional principal components (FPC) analysis, in which the FPC scores are framed as conditional expectations. This method extends the applicability of FPC analysis to longitudinal data analysis, where only a few repeated and irregularly spaced measurements are available for each subject. In PACE, the longitudinal data is modeled as noisy sampled points from a collection of trajectories that are assumed to be independent realizations of a smooth random function $X(t)$, with unknown mean function $EX(t) = \mu(t)$ and covariance function $\text{cov}(X(s), X(t)) = G(s,t)$, where $t, s \in \mathcal{T}$ and $\mathcal{T}$ is a bounded and closed time interval. The covariance function G of the process has an orthogonal expansion (in the L^2 sense) in terms of eigenfunctions ϕ_k and non-increasing eigenvalues $\lambda_k : G(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t),\ t,\ s \in \mathcal{T}$. From classical functional principal component analysis, the ith underlying random curve can be expressed through the $Karhunen - Loève$ expansion as

$$X_i(t) = \mu(t) + \sum_k \xi_{ik}\phi_k(t),\ t \in \mathcal{T} \tag{2.2.1}$$

where ξ_{ik} are uncorrelated random variables with mean 0 and variance $E\xi_{ik}^2 = \lambda_k$, with $\sum_k \lambda_k < \infty, \lambda_1 \geq \lambda_2 \geq \cdots$.

Let Y_{ij} be the jth observation of the random function $X_i(\cdot)$, observed at a random time

$T_{ij} \in \mathcal{T}$. Then Y_{ij} can be modeled as:

$$\begin{aligned} Y_{ij} &= X_i(T_{ij}) + \epsilon_{ij} \\ &= \mu(T_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(T_{ij}) + \epsilon_{ij} \end{aligned} \tag{2.2.2}$$

where ϵ_{ij} are the additional measurement errors that are assumed to be independent, identically distributed and independent of ξ_{ij}, $i = 1, ..., n$, $j = 1, ..., N_i$, N_i is the number of observations for the ith subject, and $E(\epsilon_{ij}) = 0$, $\text{var}(\epsilon_{ij}) = \sigma^2$. To reflect sparse and irregular designs, N_i are assumed to be independent and identically distributed random variables as well as independent of all other random variables.

Now we need to estimate the mean function $\mu(t)$, covariance function $G(s,t)$, eigenfunctions $\phi_k(t)$ and eigenvalues λ_k as well as functional principal component scores (FPC scores) ξ_{ik} for $k = 1, 2, ...$ for each subject $i = 1, 2, ..., n$.

Assume that the mean function, covariance function and eigenfunctions are smooth, they could be estimated by local linear smoothing. Firstly the mean function $\mu(t)$ is estimated based on the pooled data from all individuals by minimizing:

$$\sum_{i=1}^{n}\sum_{j=1}^{N_i} \kappa_1(\frac{T_{ij} - t}{h_\mu})(Y_{ij} - \beta_0 - \beta_1(t - T_{ij}))^2$$

with respect to β_0 and β_1, where $\kappa_1(\cdot)$ is kernel function and h_μ is bandwidth. Then the estimate of $\mu(t)$ is $\hat{\mu}(t) = \hat{\beta}_0(t)$.

Let $G_i(T_{ij}, T_{il}) = (Y_{ij} - \hat{\mu}(T_{ij}))(Y_{il} - \hat{\mu}(T_{il}))$ be the "raw" covariances. The local linear

surface smoother for $G(s,t)$ is defined by minimizing

$$\sum_{i=1}^{n} \sum_{1\le j\neq l\le N_i} \kappa_2(\frac{T_{ij}-s}{h_G}, \frac{T_{il}-t}{h_G}) \times (G_i(T_{ij},T_{il}) - f(\boldsymbol{\beta},(s,t),(T_{ij},T_{il})))^2,$$

where $f(\boldsymbol{\beta},(s,t),(T_{ij},T_{il})) = \beta_0 + \beta_{11}(s-T_{ij}) + \beta_{12}(t-T_{il})$, $\kappa_2(\cdot,\cdot)$ is kernel function and h_G is bandwidth. Minimization is with respect to $\boldsymbol{\beta} = (\beta_0, \beta_{11}, \beta_{12})$. Then the estimate of $G(s,t)$ is $\hat{G}(s,t) = \hat{\beta}_0(s,t)$.

The estimates of eigenfunctions $\hat{\phi}_k(t)$ and eigenvalues $\hat{\lambda}_k$ are the solutions of the eigenequations,

$$\int_{\mathcal{T}} \hat{G}(s,t)\hat{\phi}_k(s)ds = \hat{\lambda}_k\hat{\phi}_k(t).$$

The eigenfunctions can be estimated by discretizing the smoothed covariance.

Traditionally, when the measurements for each subject are densely sampled, the FPC scores $\xi_{ik} = \int (X_i(t)-\mu(t))\phi_k(t)dt$ are estimated by numerical integration. However, for longitudinal data, the time points of measurements vary widely across subjects and are sparse. The FPC scores cannot be well approximated by the usual integration method. However, as PACE further assumes that in (2.2.2), ξ_{ij} and ϵ_{ij} are jointly Gaussian, FPC scores for the ith subject could be estimated by the best prediction:

$$\tilde{\xi}_{ik} = E[\xi_{ik}|\vec{Y}_i] = \lambda_k \boldsymbol{\phi}_{ik}^T \Sigma_{Y_i}^{-1}(\vec{Y}_i - \mu_i), \tag{2.2.3}$$

where $\vec{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iN_i})^T$, $\boldsymbol{\phi}_{ik} = (\phi_k(T_{i1}), ..., \phi_k(T_{iN_i}))^T$, $\Sigma_{Y_i} = cov(\vec{Y}_i, \vec{Y}_i)$, i.e. the (j,l)'s entry of Σ_{Y_i} is $(\Sigma_{Y_i})_{j,l} = G(T_{ij},T_{il}) + \sigma^2\delta_{jl}$ with $\delta_{jl} = 1$ if $j = l$ and 0 if $j \neq l$. Substituting the parameters in (2.2.3) by the estimates of $\mu_i, \lambda_i, \phi_{ik}, \Sigma_{Y_i}$, we have the

estimation for FPC scores:

$$\hat{\xi}_{ik} = \hat{E}[\xi_{ik}|\vec{Y}_i] = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \hat{\Sigma}_{Y_i}^{-1} (\vec{Y}_i - \hat{\mu}_i), \tag{2.2.4}$$

Where the (j,l)th element of $\hat{\Sigma}_{Y_i}$ is $\hat{\Sigma}_{j,k} = \hat{G}(T_{ij}, T_{il}) + \hat{\sigma}^2 \delta_{jl}$.

Assume the infinite-dimensional processes of (2.2.1) could be approximated by the projection on the function space spanned by the first K eigenfunctions. The prediction for the trajectory $X_i(t)$ for the ith subject, using the first K eigenfunctions is:

$$\hat{X}_i^K(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \hat{\xi}_{ik} \hat{\boldsymbol{\phi}}_k(t) \tag{2.2.5}$$

According to [56], the number of eigenfunctions K can be selected by cross validation based on prediction error or by AIC-criteria. In this study we selected K by setting a threshold. That is, we selected the first K eigen functions such that they explained more than 95% of the total variation.

In summary, to recover a trajectory for each subject from longitudinal observations, we first estimate the mean curve $\hat{\mu}(t)$ and covariance $\hat{G}(s,t)$ from pooled data of all individuals, from which, eigenvalues $\hat{\lambda}_k$ and eigenfunctions $\hat{\phi}_k$ can be estimated. Then FPC scores are estimated using available observations from each subject by conditional expectation, even if only one observation is available. From (2.2.5), the FPC scores $\hat{\xi}_{ik}$ $(k = 1, \cdots, K)$ characterize each subject $i = 1, ..., n$ and can be used to describe differences between subjects. As a result, we can use FPC scores for classification or prediction [41].

For details of PACE methodology, please refer to Yao *et al.* (2015) [56]. The MATLAB package for "PACE" is available from `http://www.stat.ucdavis.edu/PACE/`.

2.2.2 Early prediction and choosing trajectory

Hall *et al.* (2012) [24] introduced a methodology for classifying and predicting the future state using functional data. It provided an approach to determine when an early decision can be made reasonably, using only part of the trajectory and showed how to use the method to choose a better biomarker as a predictor.

Hall *et al.* (2012) [24] assume there are q types of biomarkers from two classes of the population. Let $X_{ji}^{[k]}(t)$ be the observed data function of the kth biomarker from the ith subject in population Π_j, where $j = 1, 2$, $i = 1, 2, ..., n_j$, $k = 1, 2, ..., q$, $t \in \mathcal{I}$. Without loss of generality, assume $\mathcal{I} = [0, 1]$. First, the dimension of the functional data is reduced by discretizing on a grid, i.e. by confining attention to $X_{ji}^{[k]}(t_l)$, where $l = 1, 2, ..., p$ and p denotes the number of grid points. Second, a classifier based on p-variate is constructed (linear discriminant analysis and logistic classification were considered in [24]). Then the classifier is applied to each type of biomarkers using only a portion of the trajectory, for example, on the interval $\mathcal{I} = [0, t]$ with $t \in [0, 1]$, to predict which class the subject belongs to in the end, i.e. at $t = 1$. The estimated error rates are denoted by $\hat{err}^{[k]}(t)$. Comparing $\hat{err}^{[k]}(t)$ for a range of values t of interest ($t \in [0, 1]$) and $k = 1, ..., q$ gives an idea of when a relative early prediction could be made and which is a more reliable biomarker. The results are examined both numerically and theoretically by checking the consistency of $\hat{err}^{[k]}(t)$.

In this paper, we use longitudinal data for prediction, as opposed to [24] which used dense functional data. As a result, instead of reducing dimension by discretizing on a grid, we reduce dimension by PACE, which was introduced in section 2.2.1. There are two main steps in this process. First, we apply PACE to the data to calculate FPC scores for each subject. To check the appropriate time period for early prediction, using the idea from [24],

we use only part of the observations from longitudinal data in this step. Second, we apply the logistic classification method to the resulting FPC scores for prediction.

In the numerical study, we compared the one year, two years, and three years early prediction results. We also compared the error rates for different ROIs to choose which one is the most reliable biomarker.

2.2.3 Prediction with logistic classifier

Logistic regression is a special case of generalized linear models which can be used for classification. In the generalized linear model settings, logistic regression has a response Y with binomial distribution and a logit link function $logit(p) = log\frac{p}{1-p}$, where p is the probability of $Y = 1$. As in the framework of classification, the response Y denotes the index of two groups, say G_1 and G_0 with $Y = 1$ if the observation comes from class G_1 and $Y = 0$ if it comes from class G_0. Here we set $Y = 1$ for MCI-c subjects who convert from MCI to AD and $Y = 0$ for MCI-nc subjects who stay as MCI.

Suppose $x_{i1}, x_{i2}, ..., x_{iq}$ are the predictors for the subject i. The logistic regression equation is

$$logit(p_i) = log\frac{p_i}{1-p_i} = \sum_{k=1}^{q} x_{ik}\beta_k, \ \ for \ \ \ i = 1, 2, ..., n$$

where n is the number of subjects and p_i is the probability that the ith subject belongs to class G_1. The regression coefficients β_k are usually estimated using maximum likelihood estimation(MLE). Then p_i is derived by

$$p_i = \frac{1}{1 + e^{-X_i\beta}} \tag{2.2.6}$$

where $X_i = (1, x_{i1}, x_{i2}, ..., x_{iq})$, and $\beta = (\beta_0, \beta_1, \beta_2, ..., \beta_q)'$. If p_i is greater than a certain threshold (usually the threshold is set to be 0.5), subject i will be classified into class G_1, otherwise, classified into class G_0.

In this paper, we would use the trajectory of ROIs' volume as the predictor (we use function $X(t), t \in [0, T]$ to denote the trajectory). As introduced in the previous section, we firstly use PACE to reduce dimension and denote the predicted trajectory for subject i as

$$\hat{X}_i(t) = \hat{\xi}_{i1}\hat{\phi}_1(t) + \hat{\xi}_{i2}\hat{\phi}_1(t) + ... + \hat{\xi}_{iK}\hat{\phi}_1(t).$$

Thus we can do logistic regression with FPCSs $\hat{\xi}_{i1}, ..., \hat{\xi}_{iK}$ as predictors. For example, if two FPC scores are chosen, i.e. K=2, the logistic regression is

$$logit(p_i) = \beta_0 + \beta_1\hat{\xi}_{i1} + \beta_2\hat{\xi}_{i2},\ i = 1, 2, ..., n \tag{2.2.7}$$

We can easily include more variables in the logistic regression model. In the numerical study in section 2.4, in addition to FPCSs of certain ROIs, we also included some basic but important clinical features in the logistic regression model: baseline age, baseline MMSE (Mini-Mental State Examination) and APOE (apolipoprotein E) genotype. The corresponding logistical regression model is:

$$logit(p_i) = \beta_0 + \beta_1\hat{\xi}_{i1} + \beta_2\hat{\xi}_{i2} + \beta_3 Age_i + \beta_4 APOE_i + \beta_5 MMSE_i,\ i = 1, 2, ..., n \tag{2.2.8}$$

2.3 Data Description

The longitudinal volume of H, EC, MTC, FG and WB are considered to be potential predictors of conversion from MCI to AD in this paper. Baseline and follow-up volumetric MRI data, if available, were downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI, http://www.adni-info.org/) database along with the corresponding clinical information. The structural MRI scans were acquired from 1.5T or 3T scanners, manufactured by GE, Siemens or Philips. Regional brain segmentation and volume estimation were carried with FreeSurfer by the UCSF/SF VA Medical Center [25], [49].

From a total of 872 individuals with a baseline diagnosis of MCI who were recruited for ADNI, 66 subjects were excluded due to missing data and 5 subjects were excluded due to reverting from AD to MCI. In the end 801 MCI subjects were included in this study. They were split into two categories. The MCI subjects who converted to AD after some years are labeled as MCI-c (mild cognitive impairment converters; $n = 272$); the rest of them who did not convert to AD during the follow-up period were labeled as MCI-nc (mild cognitive impairment nonconverters; $n = 529$). At study entry (baseline), all subjects underwent a comprehensive clinical evaluation, cognitive/functional assessments, and a structural brain MRI scan. Subjects also provided a blood sample for apolipoprotein E (APOE) genotyping and proteomic analysis. The subjects were then followed longitudinally at specific time points (6, 12, 18, 24, 36... months). However, due to the variability of their visits, the data collected is irregular in the number of observations and in time intervals for each subject. Table 2.1 shows the characteristics of the MCI subjects included in this study. It shows, except for gender, the other three features (baseline age, baseline MMSE, and APOE) are significantly different between MCI-c and MCI-nc subjects at a significance level of 0.05,

Table 2.1 Subjects characteristics

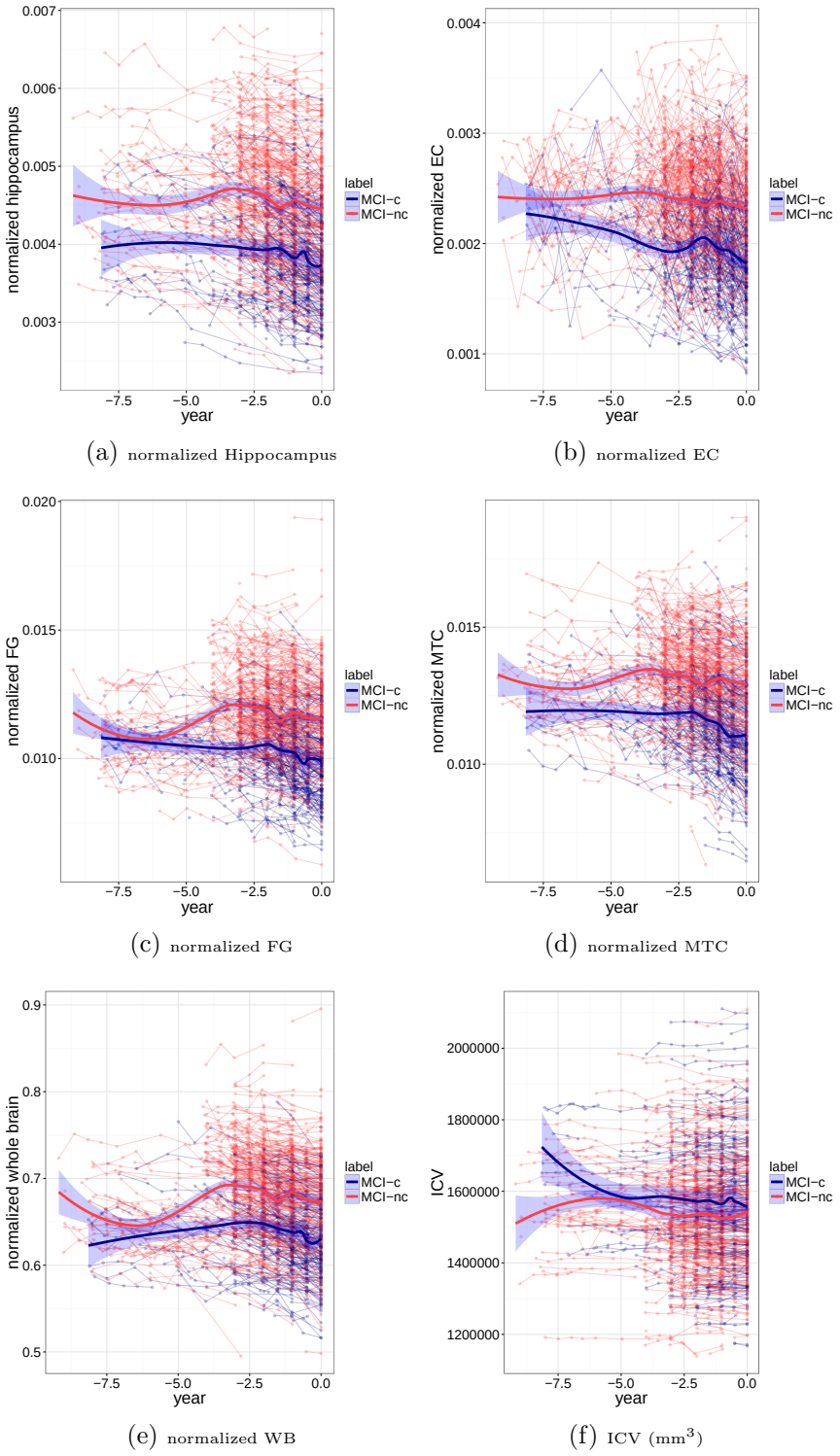
	MCI-c	MCI-nc	p-value
n	272	529	
Age (Mean±sd)	73.5 ± 7.03	72.3 ± 7.59	0.019
Gender (F/M)	110/162	220/309	0.813
MMSE (Mean±sd)	26.9 ± 1.77	28.0 ± 1.70	< 0.001
APOE (+/-)	182/90	226/303	< 0.001

Key: MCI-c, mild cognitive impairment (converters); MCI-nc, mild cognitive impairment (non converters); Age, baseline age; MMSE, baseline Mini-Mental State Examination; APOE, apolipoprotein E genotype .

which suggests the three features might be helpful in prediction. For all the subjects, the longest observed year is about 9.18 years and the minimum observed year is 0 (only one data point is collected, see Table 2.2). The numbers of observations for the subjects are also summarized in Table 2.2. The maximum number of observations is 10, the minimum number is 1, and the median number is about $3 - 5$ in each category.

During the data analysis, all regional volumes were normalized by dividing the intracranial volume (ICV) to correct for individual differences in head size. For example, to normalize the hippocampal volume, we calculate the fraction of the hippocampal volume and the ICV from the same subject at the same time point. The resulting fractions are then used in the analysis. In the spaghetti plot (Figure 2.1), (f) is the longitudinal ICV for each subject. It shows that the ICV for MIC-c and MIC-nc subjects have a similar pattern on average. Also for each subject, as expected the ICV is relatively constant during the observed period. For the other five normalized regional volumes ((a)-(e) in Figure 2.1), on average the MCI-c group has a higher rate of atrophy in volume compared to MCI-nc group.

As shown in the spaghetti plot (Figure 2.1), first, we aligned the longitudinal volumetric data of ROIs by the timeline. Specifically, the MCI-c subjects were aligned by "time of conversion", i.e. the time point of conversion is defined as year 0 and year -1 is defined for



(a) normalized Hippocampus (b) normalized EC

(c) normalized FG (d) normalized MTC

(e) normalized WB (f) ICV (mm^3)

Figure 2.1 Longitudinal volume of ROIs for MCI subjects. In (a)-(e), the value on Y axis is the normalized volume. In (f), the value on Y axis is the ICV (mm^3). The value on X axis is "disease year". Thin lines are observations for each subject. Blue lines are for MCI-c subjects and red lines are for MCI-nc subjects. Blue and red thick lines are pooled mean curves for MCI-c group and MCI-nc group respectivly.

Table 2.2 Data characteristics

	n	year (max)	year (min)	year (median)	obsN (max)	obsN (min)	obsN (median)
All subjects							
MCI-c	272	8.15	0	1.65	8	1	3
MCI-nc	529	9.18	0	2.07	10	1	3
All 3y subjects							
MCI-c	30	8.10	2.97	3.03	8	4	5
MCI-nc	97	8.03	2.52	3.02	10	3	4

Key: MCI-c, mild cognitive impairment (converters); MCI-nc, mild cognitive impairment (non converters). "All 3y subjects" are the subjects who have observations at all the "-3y", "-2y", "-1y" time points. Column 2 is the number of subjects in each category. Column $3-5$ are the maximum, minimum and median length of observed years. Column $6-8$ are the maximum, minimum and median number of observations.

the time points observed 1 year prior to conversion. The MCI-nc subjects were aligned by the last time point of observation, i.e. the time point of last observation is defined as year 0 and the time points observed 1 year prior to the last observation is defined as year "-1" and so forth. In this way, the data were homogenized by the timeline of disease progression. The goal is to predict whether a subject will convert to AD in the end, i.e. at year 0.

One of the goals of this paper is also to determine a reasonable prediction window. So we need to compare the prediction accuracy among different prediction windows. Here we compare one year, two years and three years of early prediction. To make the three years' prediction results comparable, we pick the subjects who have observations at all the three most recent years: year "-3", year "-2", year "-1" as the testing set (labeled as "all 3y" subjects in Table 2.2 and 2.3). There are 127 "all 3y" subjects in total, of which 30 are MCI-cs and 97 are MCI-ncs. In Table 2.3, characteristics of the subjects who have observations at year "-3", "-2", and "-1" are listed in the first three columns respectively. The last column in Table 2.3 lists the characteristics of the testing set. All the subjects are considered as the training set. Leave-one-out testing is applied to the subjects in the testing set.

Table 2.3 Subjects characteristics by year

	-3y	-2y	-1y	all 3y
MCI-c				
n	71	158	245	30
Age (mean±SD)	73.94±7.06	73.61±7.05	73.47±7.20	73.94±5.98
MMSE (mean±SD)	27.14±1.82	27.05±1.69	26.94±1.78	27.03±1.63
Gender (F/M)	28/43	58/100	100/145	10/20
APOE (+/-)	40/31	103/55	162/83	19/11
MCI-nc				
n	242	371	473	97
Age (mean±SD)	72.09±7.61	72.43±7.36	72.39±7.47	70.22±7.20
MMSE (mean±SD)	28.03±1.64	28.08±1.61	27.97±1.7	28.12 ±1.62
Gender (F/M)	99/143	154/217	197/276	40/57
APOE (+/-)	90/152	148/223	203/270	40/57

Key: MCI-c, mild cognitive impairment (converters); MCI-nc, mild cognitive impairment (non converters); MMSE, baseline Mini-Mental State Examination; APOE, apolipoprotein E genotype."-3y" column, "-2y" column and "-1y" column are characteristics of subjects who have observations at year "-3", "-2", and "-1" respectively after alignment. "all 3y" column is the characteristics of the subjects who have observations at all the 3 years: year "-3", "-2" and "-1".

2.4 Numerical results

This section presents the numerical results and conclusions. We first took the hippocampus as an example to illustrate the details of the prediction procedure using a single ROI's volume in section 2.4.1. We then list all prediction results using a single ROI in section 2.4.2. In the end, we investigate whether or not any combinations of the ROIs would improve the prediction performance in section 2.4.3. There are 26 different combinations of the 5 ROIs. The prediction results from the 26 combination prediction models are listed in this section.

2.4.1 Prediction using hippocampus

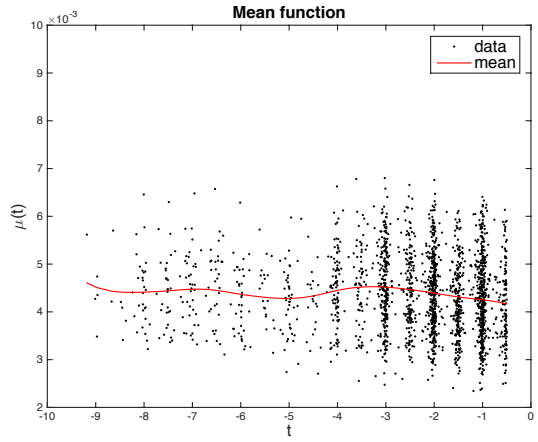
In this section, we take the volume of the hippocampus as an example to describe how to use a single ROI for prediction. We started with the 1-year early prediction.

Table 2.4 Parameter estimation (hippocampus for 1y prediction)

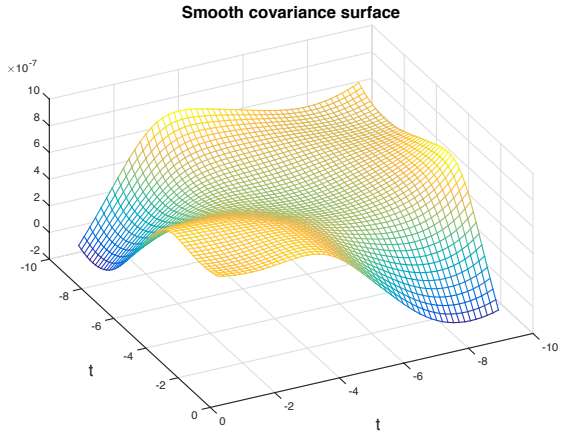
	Estimate	Std. Error	z value	p-value
(Intercept)	6.597	1.779	3.71	2.07e-04
APOE4	0.8120	0.1802	4.51	6.62e-06
Age	-0.0225	0.0132	-1.71	8.70e-02
MMSE	-0.2243	0.0500	-4.48	7.41e-06
FPC Score1	5265	1631	3.23	1.25e-03
FPC Score2	20364	5776	3.53	4.22e-04

The first step is PACE analysis with the partially observed data and calculating the FPC scores for each subject. As outlined in section 2.3, the longitudinal data were realigned according to disease status (Figure 2.1). We wanted to predict whether or not a subject would convert to AD at year 0. 1-year prediction means to use all the data observed prior to year "-1" to predict the state at year 0. Since the data points were irregularly collected, we use the data with year $t \in [-9.18, -0.5)$ for 1-year prediction (9.18 is the longest trajectory for all the subjects and all ROIs, see Figure 2.1). Figure 2.2 shows the PACE analysis results for 1-year early prediction using hippocampus. (a) and (b) are the estimation of mean curve $\mu(t)$ and covariance surface $G(s,t)$ for $s,t \in [-9.18, -0.5)$, which are introduced in section 2.2.1. We selected the number of the eigenfunctions used in (2.2.5) by setting a threshold to be 95%. (d) shows that the first two principal components explained 98.9% of the variance. Thus only the first two eigenfunctions and corresponding FPC scores $\hat{\xi}_{i1}$ and $\hat{\xi}_{i2}$ of each subject i are employed for prediction. (c) is the estimation of the first two eigenfunctions $\phi_1(t)$ and $\phi_2(t)$ for $t \in [-9.18, -0.5)$. The FPC scores are calculated by (2.2.4) and then applied to logistic regression model (2.2.8). Figure 2.2.3 shows the scatter plot of the first two FPC scores for all the training subjects.

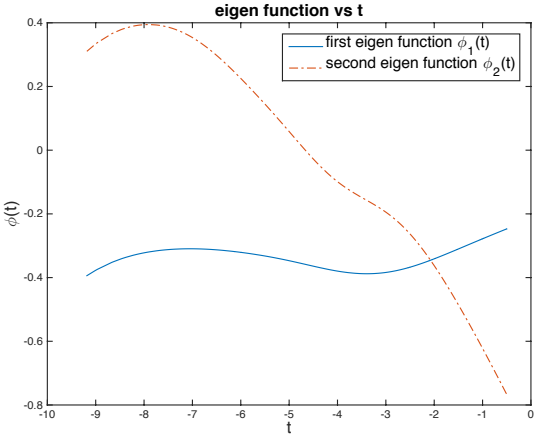
Table 2.4 provides sample statistics for the coefficient estimates which are obtained from the logistic regression of group indicator ("MCI-c"= 1, "MCI-nc"= 0) with linear model



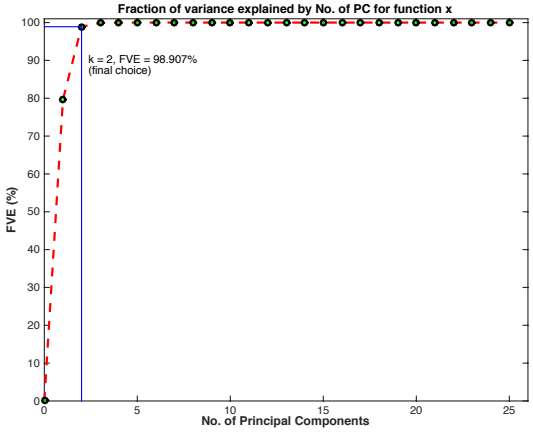
(a) Mean curve $\hat{\mu}(t)$ of hippocampus



(b) Covariance surface $\hat{G}(s,t)$ of hippocampus



(c) The first two eigen function



(d) Fraction of variance explained by eigen function

Figure 2.2 PACE analysis using hippocampus for 1-year prediction. (a)-(c) show the estimations of mean function $\mu(t)$, covariance surface $G(t)$ and the first two eigen functions $\phi_1(t)$ and $\phi_2(t)$. (d) shows the first two eigen functions explained 98.907% of the total variance.

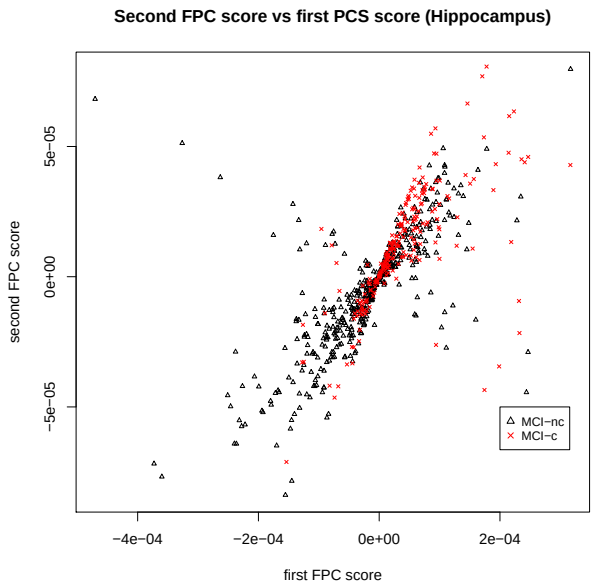


Figure 2.3 Second versus first FPC scores for hippocampus (1 year prediction). The triagulars indicate MCI-nc and the crosses indicate MCI-c.

(2.2.8). It shows both the first two FPC scores $\hat{\xi}_1$ and $\hat{\xi}_2$ are significant in increasing the odds of being MCI-c. The resulting estimates of coefficients from the logistic models (2.2.8) are plugged into (2.2.6) to calculate the conversion probability. We let p_i be the probability of conversion to AD for subject i. Let p_0 be a threshold. If $p_i \geq p_0$, subject i will be classified as MCI-c, otherwise, it will be classified as MCI-nc. If set $p_0 = 0.5$, the sensitivity, specificity, and accuracy for 1-year early prediction are $0.7, 0.86$ and 0.82. By changing the threshold p_0 from 0 to 1, we can derive a ROC (receiver operating characteristic) curve with AUC (area under the curve) to be 0.838.

We then apply the same procedure mentioned above to a 2-year early prediction and a 3-year early prediction with hippocampus volume. Setting threshold p_0 to be 0.5, the corresponding sensitivity, specificity, and accuracy are summarized in Table 2.5. Setting the threshold p_0 to be from 0 to 1, the ROC curve are shown in Figure 2.4. From the ROC curve, shorter prediction windows perform better, i.e., 1-year early prediction is the best and 3-year

Table 2.5 Prediction accuracy rates (hippocampus)

	-1 y	-2 y	-3y
sensitivity	0.7	0.53	0.23
specificity	0.86	0.89	0.98
accuracy	0.82	0.80	0.80

Key: The probability threshold p_0 is set to be 0.5, i.e. if $p \geq 0.5$, classify it as MCI-c, otherwise classify it as MCI-nc.

early prediction is the worst. It is expected, and the procedure provides risk evaluation in early prediction.

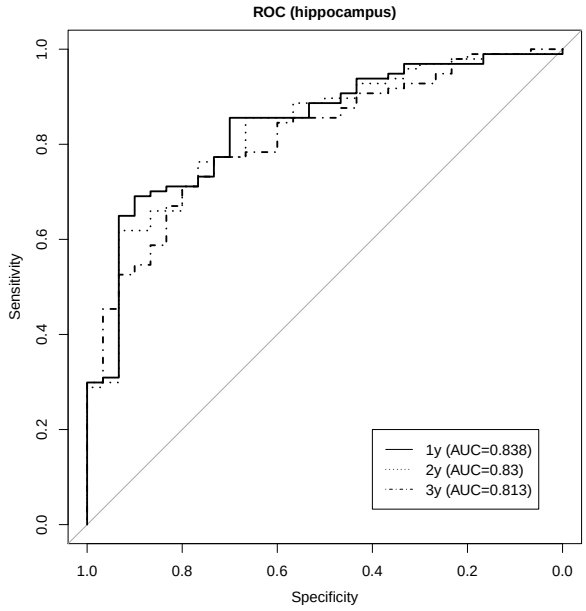


Figure 2.4 ROC curve for prediction using hippucampus. Solid line, dotted line and dotdash line are the ROC curves for 1-year, 2-year and 3-year prediction respectively.

2.4.2 Prediction results using single ROI

In section 2.4.1, we took the hippocampus as an example to illustrate the procedure of 1-year, 2-year and 3-year early prediction using a single ROI. The results shown in section 2.4.1 were based on one training data set. In this section, we apply the same procedure to all the five ROIs: H, WB, EC, FG and MTC. To check the robustness of prediction performance,

we repeated the procedure on 100 samples of training data. That is, we randomly sampled 2/3 of the training data (2/3 MCI-c and 2/3 MCI-nc) for 100 times. We then repeated the prediction procedures stated in section 2.4.1 for 100 times. The mean and standard deviation of sensitivity, specificity, and accuracy were calculated when classification probability threshold p_0 was set to be 0.5. Also, when set p_0 to vary from 0 to 1, the mean and standard deviation of the AUCs from the ROC curves were also calculated. The results are shown in Table 2.6.

From Table 2.6, all the five ROIs have prediction accuracy above 75% and AUC around 0.8 for all 1-year, 2-year, and 3-year prediction. Both H and EC can identify MCI-c correctly with a greater-than-chance probability as early as two years. EC is the best ROI with prediction accuracy above 80% and AUC above 0.8 for all the three years' predictions. Moreover, its 2-year sensitivity is 62% and 1-year sensitivity achieves 70%.

2.4.3 Prediction results using combinations of ROIs

We then check the prediction performance from the combinations of ROIs. Combination prediction is realized by plug combinations of ROIs' FPC scores into the logistic regression model. For example, if we predict using the combination of H and EC, we first calculate the FPC scores of H ($\hat{\xi}_{1i}, \hat{\xi}_{2i}$) as well as FPC scores of EC ($\hat{\eta}_{1i}, \hat{\eta}_{2i}$) for each subject i (suppose the number of the eigenfunctions selected for both H and EC is 2). Then the logistic regression model for predicting using combination of H and EC is:

$$logit(p_i) = \alpha + \beta_1\hat{\xi}_{i1} + \beta_2\hat{\xi}_{i2} + \beta_3\hat{\eta}_{i1} + \beta_4\hat{\eta}_{i2} + \beta_5 Age_i + \beta_6 APOE_i + \beta_7 MMSE_i,$$

for $i = 1, 2, ..., n$.

Table 2.6 Prediction performance using single ROI

		sensitivity	specificity	accuracy	AUC
H	1 y	0.64±0.044	0.85±0.013	0.8±0.013	0.84±0.003
(Mean ± SD)	2 y	0.51±0.029	0.89±0.006	0.8±0.007	0.83±0.004
	3 y	0.27±0.054	0.96±0.018	0.8±0.011	0.81±0.006
WB	1 y	0.54±0.039	0.86±0.015	0.79±0.012	0.8±0.004
(Mean ± SD)	2 y	0.37±0.051	0.88±0.013	0.76±0.01	0.8±0.004
	3 y	0.1±0.027	0.95±0.015	0.75±0.01	0.78±0.005
EC	1 y	0.7±0.011	0.85±0.016	0.82±0.012	0.84±0.003
(Mean ± SD)	2 y	0.62±0.03	0.91±0.012	0.84±0.009	0.84±0.004
	3 y	0.32±0.037	0.95±0.01	0.8±0.009	0.81±0.004
FG	1 y	0.53±0.03	0.88±0.006	0.8±0.006	0.82±0.003
(Mean ± SD)	2 y	0.46±0.026	0.89±0.008	0.79±0.006	0.81±0.003
	3 y	0.23±0.039	0.95±0.015	0.78±0.01	0.79±0.003
MTC	1 y	0.57±0.043	0.88±0.015	0.8±0.011	0.84±0.005
(Mean ± SD)	2 y	0.42±0.027	0.9±0.011	0.79±0.01	0.83±0.006
	3 y	0.24±0.038	0.94±0.012	0.78±0.01	0.81±0.007

Key: H stands for hippocampus, WB stands for whole brain, EC stands for entorhinal cortex, FG stands for fusiform gyrus and MTC stands for middle temporal cortex. Sensitivity is the proportion of true MCI-c in prediction, specificity is the proportion of true MCI-nc in prediction and accuracy is the proportion of true prediction. p_0 is set to be 0.5. AUC is the area under the curve from the ROC (receiver operatin characteristic) curve. The mean and standard deviation of sensitivity, specificity, accuracy, and AUC are based on 100 samples of training data sets.

There are 26 different combinations in total. Table 2.7 shows the results of combinations of two ROIs. Table 2.8 shows the results of combinations of three ROIs. Table 2.9 shows the results of combinations of more than three ROIs. Please see Table 2.7, Table 2.8 and Table 2.9 in Appendix.

2.4.4 Conclusion

The graph comparison of the prediction performances from the 31 different combinations of the ROIs (5 single variable models and 26 combination models) are listed in Figure 2.5. From Table 2.6, 2.7, 2.8, 2.9 and Figure 2.5, we have the following observations:

First, the longitudinal volumes from the listed brain ROIs measured by MRI have prediction power as early as three years in advance. In 1-year, 2-year, and 3-year prediction for all the models, the overall specificity is above 80%. The accuracy is above 70%, and the AUC is above 0.8. The sensitivity varies for different prediction windows and models from 10% to 70%.

Second, short-term prediction performs better, i.e. 1-year early prediction performs the best compare to 2-year and 3-year prediction. It is consistent with intuition because more information is included and less noise is introduced in short-term prediction procedure. But the error curves over time are clearly helpful for early interventions.

To compare the models using different combinations of ROIs, since the overall specificity, accuracy, and AUC are similar among models, we mainly focus on checking the sensitivity in all models. Sensitivity is the probability that the MCI-c subjects are correctly identified. It is crucial in prognosis and clinical trials. Overall, most of the models derived a sensitivity of greater than 50% for 1-year and 2-year prediction. However, the maximum sensitivity for 3-year in advance prediction is just about 35%, which implies it is not easy to correctly

identify MCI-c subjects three years in advance using the information in our model.

Among the five models using single ROI, EC performs the best, followed by H, and WB is the worst. All the five ROIs have sensitivity higher than 50% for 1-year prediction. Both H and EC have a sensitivity greater than 50% in 2-year early prediction (Table 2.6). 3-year prediction sensitivity are about 10% to 30% for all the ROIs.

For the prediction results from the models using combinations of two or more ROIs (see Table 2.7), except for the combinations of WB+FG and WB+MTC , all the other ROIs' combinations have a sensitivity higher than chance for both 1-year and 2-year prediction. The overall sensitivity for 3-year prediction is about 20%-30% for models of combinations of two ROIs, while which is around 30% for models of three ROIs and around 35% for models of more than three ROIs (see Table 2.7 and Table 2.8).

To select the best overall prediction performance in the sense of prediction window, sensitivity, specificity, accuracy, and AUC, the results from the models of EC (see Table 2.6), H+EC+FG (see Table 2.8), H+WB+EC+FG and H+WB+EC+FG+MTC (see Table 2.9) are the best. They all have the highest level of prediction sensitivity, specificity, accuracy, and AUC. The highest 1-year sensitivity is about 70%, specificity is about 86%, accuracy is about 82%, and AUC is about 0.85. For 2-year prediction, they also have the highest prediction rate with sensitivity as about $62\% - 65\%$, specificity as about 90%, accuracy as about 84%, and AUC as about 0.85. For 3-year early prediction, they all have sensitivity above 32%, specificity above 90%, accuracy above 80%, and AUC above 0.81.

In the four best models, the combination models work a little bit better than single ROI model. Because the three combination models have a higher AUC than that of EC model. However, they are not significantly superior than EC. On the other hand, the combinations without EC and H, which are the first and second best single biomarker, perform relatively

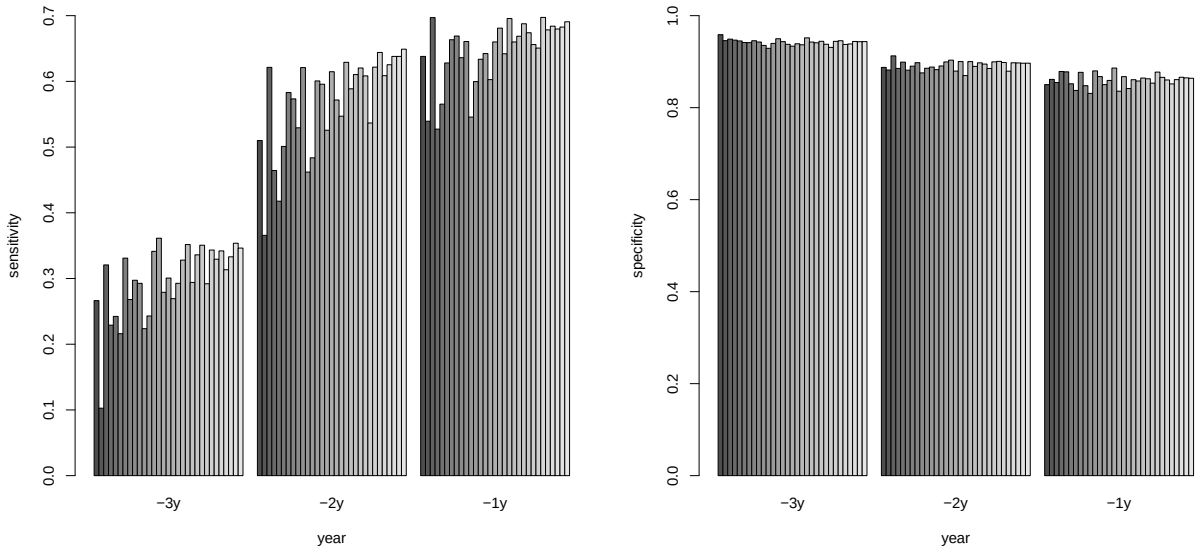
worse than other combinations (WB+FG, WB+MTC, FG+MTC, WB+FG+MTC, see Table 2.7 and Table 2.8). This observation implies that the prediction power of the combination models comes from the prediction power of EC and H.

2.5 Discussion

In this study we use sparsely observed volumes of ROIs quantified by longitudinal volumetric MRI to predict whether or not a MCI subject will convert to AD in a specified time window. A longitudinal data prediction method based on functional data analysis is developed for early prediction in varying time windows, as well as identifying the most important ROI(s) in the process.

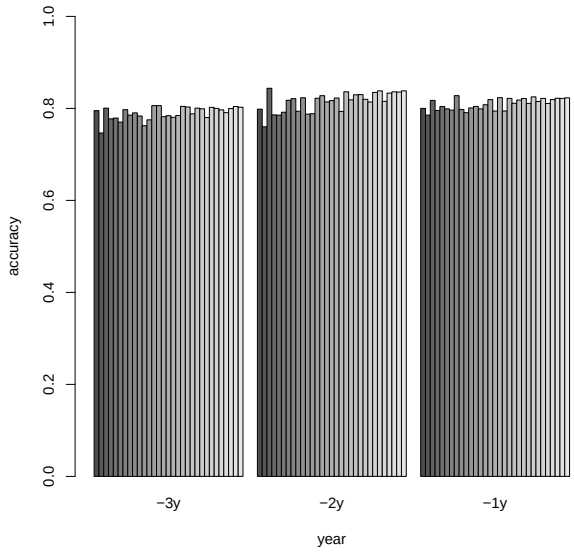
The application of existing methods is not obvious due to the complexity of the data structure. To apply the prediction method based on functional data analysis, we first use PACE to extract statistically validated information from the longitudinal data. FPC scores are calculated and used for prediction. This method is new for analyzing ADNI data.

Our best prediction model (1-year prediction using H+WB+EC+FG+MTC with AUC = 0.86, accuracy = 82%, specificity=86%, and sensitivity=69%) performs favorably compared to the existing literature for prediction using the same longitudinal data, even though the comparison is not entirely comparable since those studies used different biomarkers and different prediction windows. Actually, our model used fewer predictors (longitudinal ROI volumes and three clinical features) and has a longer prediction window (one year, two years, and three years), which suggests that our model efficiently used less information and derived comparable prediction results. Since the primary goal of this paper is to introduce a statistically validated, advanced methodology for prediction using longitudinal data, we

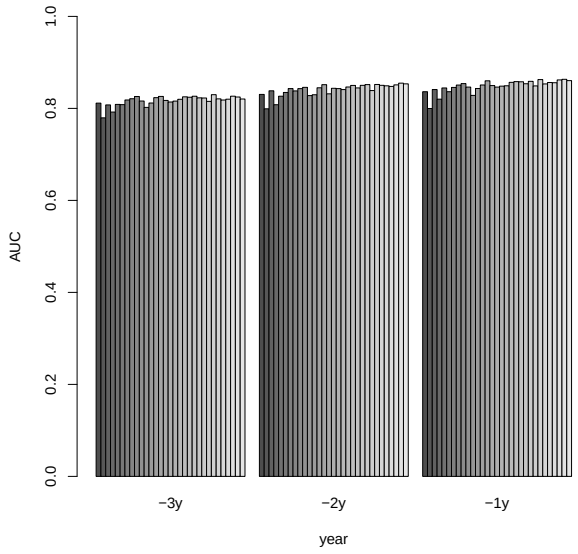


(a) sensitivity

(b) specificity



(c) accuracy



(d) AUC

Figure 2.5 Prediction sensitivity(a), specificity(b), accuracy(c), and AUC(d) using single and combinations of ROIs. In each panel, every bar represent a predict result using different combinations of biomarkers (from left to right): Hippocampus(H), whole brain(W), entorhinal cortex(EC), fusiform gyrus(F), middle temporal cortex(MTC), H+WB, H+EC, H+FG, H+MTC, WB+EC, WB+FG, WB+MTC, EC+FG, EC+MTC, FG+MTC, H+WB+EC, H+WB+FG, H+WB+MTC, H+EC+FG, H+EC+MTC, H+FG+MTC, WB+EC+FG, WB+EC+MTC, WB+FG+MTC, EG+FG+MTC, H+WB+EC+FG, H+WB+EC+MTC, H+WB+FG+MTC, H+EC+FG+MTC, WB+EC+FG+MTC and combination of all the five ROIs.

do not intend to elaborate the results by including more predictors. However, one could continue experimentation with other available biomarkers by the techniques adapted here.

Besides, due to the irregularity and sparsity of longitudinal data, most other literature make predictions just in one prediction window. Our proposed method makes it feasible for predicting on more flexible prediction windows, even if only one observation is available. It enables us to compare the prediction performance in varying prediction windows and helps to understand the disease risk over time.

Moreover, most of the methods employed in the literature are machine learning techniques, which are not accessible for statistical validation and interpretation. Our model is based on two well-established statistical methods, which is easy to validate and interpret.

One limitation of the proposed model is it has a necessary assumption for PACE method. We need to assume the measurement error ϵ_{ij} and FPC scores ξ_{ij} are jointly Gaussian. This assumption is not always valid. However, the existing simulation studies indicate the method is robust to some extent in violations of the Gaussian assumption (see [56]). Also, in the prediction result, the sensitivity is poor for all combinations of the ROIs, especially for 3-year prediction. This reveals the limitation of using volume of ROI for prediction. More advanced classification technique would be applied to improve the prediction sensitivity in the future work.

In summary, the proposed procedure in this paper provides a method to predict conversion from MCI to AD using longitudinal data. The longitudinal prediction curve can be utilized for understanding disease risk over time. The procedure can be easily applied to analyze other longitudinal data sets for prediction in clinical trials to decide a better prediction window and choose better biomarker(s).

APPENDIX

Table 2.7 Prediction performance of combinations of two ROIs

		sensitivity	specificity	accuracy	AUC
H+WB	1 y	0.63±0.044	0.85±0.012	0.8±0.01	0.84±0.004
(Mean ± SD)	2 y	0.5±0.037	0.88±0.008	0.79±0.011	0.83±0.006
	3 y	0.22±0.065	0.94±0.016	0.77±0.013	0.81±0.008
H+EC	1 y	0.66±0.031	0.84±0.011	0.8±0.01	0.85±0.003
(Mean ± SD)	2 y	0.58±0.044	0.89±0.013	0.82±0.015	0.84±0.003
	3 y	0.33±0.029	0.94±0.012	0.8±0.009	0.82±0.005
H+FG	1 y	0.67±0.032	0.88±0.016	0.83±0.01	0.85±0.003
(Mean ± SD)	2 y	0.57±0.035	0.9±0.01	0.82±0.01	0.84±0.004
	3 y	0.27±0.049	0.95±0.009	0.79±0.01	0.82±0.005
H+MTC	1 y	0.64±0.039	0.85±0.014	0.8±0.011	0.85±0.004
(Mean ± SD)	2 y	0.53±0.033	0.88±0.016	0.79±0.012	0.84±0.005
	3 y	0.3±0.052	0.94±0.011	0.79±0.013	0.83±0.006
WB+EC	1 y	0.66±0.025	0.83±0.015	0.79±0.012	0.85±0.004
(Mean ± SD)	2 y	0.62±0.022	0.89±0.014	0.82±0.012	0.85±0.004
	3 y	0.29±0.026	0.94±0.015	0.78±0.009	0.82±0.005
WB+FG	1 y	0.55±0.052	0.88±0.007	0.8±0.012	0.83±0.006
(Mean ± SD)	2 y	0.46±0.026	0.89±0.006	0.79±0.006	0.83±0.007
	3 y	0.22±0.048	0.93±0.012	0.76±0.011	0.8±0.007
WB+MTC	1 y	0.6±0.054	0.87±0.014	0.8±0.012	0.84±0.005
(Mean ± SD)	2 y	0.48±0.047	0.88±0.016	0.79±0.014	0.83±0.007
	3 y	0.24±0.043	0.94±0.011	0.78±0.011	0.81±0.005
EC+FG	1 y	0.63±0.032	0.85±0.014	0.8±0.01	0.85±0.003
(Mean ± SD)	2 y	0.6±0.011	0.89±0.01	0.82±0.007	0.84±0.003
	3 y	0.34±0.038	0.95±0.009	0.81±0.009	0.82±0.003
EC+MTC	1 y	0.64±0.016	0.86±0.012	0.81±0.008	0.86±0.003
(Mean ± SD)	2 y	0.6±0.029	0.9±0.012	0.83±0.007	0.85±0.004
	3 y	0.36±0.03	0.94±0.009	0.81±0.008	0.83±0.005
FG+MTC	1 y	0.6±0.02	0.89±0.009	0.82±0.006	0.85±0.004
(Mean ± SD)	2 y	0.53±0.032	0.9±0.01	0.81±0.009	0.83±0.005
	3 y	0.28±0.055	0.94±0.01	0.78±0.012	0.82±0.005

Key: H stands for hippocampus, WB stands for whole brain, EC stands for entorhinal cortex, FG stands for fusiform gyrus and MTC for middle temporal cortex. Sensitivity is the proportion of true MCI-c in prediction, specificity is the proportion of true MCI-nc in prediction, and accuracy is the proportion of true prediction. p_0 is set to be 0.5. AUC is the area under the curve from the ROC (receiver operating characteristic) curve. The mean and standard deviation of the sensitivity, specificity, accuracy, and AUC are based on 100 samples of training data sets.

Table 2.8 Prediction performance of combinations of three ROIs

		sensitivity	specificity	accuracy	AUC
H+WB+EC	1 y	0.66±0.029	0.84±0.014	0.79±0.011	0.85±0.003
(Mean ± SD)	2 y	0.61±0.031	0.88±0.011	0.82±0.012	0.84±0.005
	3 y	0.3±0.03	0.93±0.015	0.78±0.012	0.81±0.006
H+WB+FG	1 y	0.68±0.03	0.87±0.013	0.82±0.009	0.85±0.005
(Mean ± SD)	2 y	0.57±0.055	0.9±0.011	0.82±0.016	0.84±0.007
	3 y	0.27±0.052	0.94±0.01	0.78±0.012	0.82±0.008
H+WB+MTC	1 y	0.64±0.048	0.84±0.013	0.79±0.012	0.85±0.005
(Mean ± SD)	2 y	0.55±0.045	0.87±0.013	0.79±0.011	0.84±0.006
	3 y	0.29±0.05	0.94±0.013	0.78±0.014	0.82±0.006
H+EC+FG	1 y	0.7±0.013	0.86±0.017	0.82±0.013	0.86±0.003
(Mean ± SD)	2 y	0.63±0.03	0.9±0.012	0.84±0.012	0.85±0.004
	3 y	0.33±0.037	0.95±0.009	0.8±0.01	0.83±0.004
H+EC+MTC	1 y	0.66±0.018	0.86±0.012	0.81±0.009	0.86±0.004
(Mean ± SD)	2 y	0.59±0.041	0.89±0.012	0.82±0.009	0.85±0.003
	3 y	0.35±0.03	0.94±0.01	0.8±0.008	0.82±0.006
H+FG+MTC	1 y	0.67±0.036	0.86±0.016	0.82±0.01	0.86±0.003
(Mean ± SD)	2 y	0.61±0.041	0.9±0.011	0.83±0.011	0.84±0.005
	3 y	0.29±0.05	0.94±0.008	0.79±0.012	0.83±0.005
WB+EC+FG	1 y	0.69±0.023	0.86±0.017	0.82±0.014	0.85±0.004
(Mean ± SD)	2 y	0.62±0.032	0.89±0.009	0.83±0.009	0.85±0.006
	3 y	0.34±0.047	0.94±0.009	0.8±0.012	0.82±0.006
WB+EC+MTC	1 y	0.67±0.032	0.85±0.011	0.81±0.01	0.86±0.004
(Mean ± SD)	2 y	0.61±0.029	0.89±0.012	0.82±0.01	0.85±0.005
	3 y	0.35±0.034	0.94±0.011	0.8±0.011	0.82±0.006
WB+FG+MTC	1 y	0.66±0.047	0.88±0.011	0.83±0.011	0.85±0.006
(Mean ± SD)	2 y	0.54±0.034	0.9±0.013	0.81±0.013	0.84±0.009
	3 y	0.29±0.055	0.93±0.008	0.78±0.013	0.82±0.007
EC+FG+MTC	1 y	0.65±0.029	0.87±0.016	0.82±0.01	0.86±0.003
(Mean ± SD)	2 y	0.62±0.017	0.9±0.011	0.83±0.009	0.85±0.004
	3 y	0.34±0.032	0.94±0.008	0.8±0.01	0.83±0.004

Key: H stands for hippocampus, WB stands for whole brain, EC stands for entorhinal cortex, FG stands for fusiform gyrus and MTC for middle temporal cortex. Sensitivity is the proportion of true MCI-c in prediction, specificity is the proportion of true MCI-nc in prediction, and accuracy is the proportion of true prediction. p_0 is set to be 0.5. AUC is the area under the curve from the ROC (receiver operating characteristic) curve. The mean and standard deviation of the sensitivity, specificity, accuracy, and AUC are based on 100 samples of training data sets.

Table 2.9 Prediction performance of combinations of four or more ROIs

		sensitivity	specificity	accuracy	AUC
H+WB+EC+FG	1 y	0.7±0.011	0.86±0.015	0.82±0.012	0.85±0.005
(Mean ± SD)	2 y	0.64±0.038	0.9±0.012	0.84±0.013	0.85±0.007
	3 y	0.33±0.039	0.95±0.01	0.8±0.011	0.82±0.007
H+WB+EC+MTC	1 y	0.68±0.029	0.85±0.012	0.81±0.009	0.86±0.004
(Mean ± SD)	2 y	0.61±0.036	0.88±0.013	0.82±0.011	0.85±0.005
	3 y	0.34±0.036	0.94±0.011	0.8±0.012	0.82±0.007
H+WB+FG+MTC	1 y	0.68±0.04	0.86±0.017	0.82±0.012	0.86±0.005
(Mean ± SD)	2 y	0.63±0.046	0.9±0.012	0.83±0.014	0.85±0.008
	3 y	0.31±0.054	0.94±0.008	0.79±0.013	0.82±0.007
H+EC+FG+MTC	1 y	0.68±0.022	0.87±0.014	0.82±0.011	0.86±0.003
(Mean ± SD)	2 y	0.64±0.023	0.9±0.012	0.84±0.01	0.85±0.004
	3 y	0.33±0.034	0.94±0.008	0.8±0.009	0.83±0.005
WB+EC+FG+MTC	1 y	0.68±0.031	0.86±0.013	0.82±0.011	0.86±0.004
(Mean ± SD)	2 y	0.64±0.028	0.9±0.01	0.84±0.01	0.85±0.006
	3 y	0.35±0.038	0.94±0.009	0.8±0.011	0.82±0.007
H+WB+EC+FG+MTC	1 y	0.69±0.022	0.86±0.012	0.82±0.01	0.86±0.005
(Mean ± SD)	2 y	0.65±0.033	0.9±0.012	0.84±0.012	0.85±0.007
	3 y	0.35±0.038	0.94±0.008	0.8±0.011	0.82±0.008

Key: H stands for hippocampus, WB stands for whole brain, EC stands for entorhinal cortex, FG stands for fusiform gyrus and MTC for middle temporal cortex. Sensitivity is the proportion of true MCI-c in prediction, specificity is the proportion of true MCI-nc in prediction, and accuracy is the proportion of true prediction. p_0 is set to be 0.5. AUC is the area under the curve from the ROC (receiver operating characteristic) curve. The mean and standard deviation of the sensitivity, specificity, accuracy, and AUC are based on 100 samples of training data sets.

Chapter 3

High dimensional discriminant analysis for spatially dependent data and its application in neuroimaging data

3.1 Introduction

Spatial statistics is rapidly developing for analyzing data which is featured with spatial structure. Applications of spatial statistics are for a broad range of disciplines such as agriculture, geology, soil science, oceanography, forestry, meteorology and climatology as well as imaging data.

With the development of computer technology into the processing of images, spatial statistics has played an important role in the processing of images and pattern recognition. Cressie (1992) [17] introduced details of spatial methodologies in analysis of images. Motivated by the problem of how to use brain imaging data for diagnosis and classification, we focus on discriminant analysis of data featured with spatial correlation.

Another important characteristic of the imaging data is the high dimensionality. For

example, in MRI scans, signals are collected from a 3D space. The number of voxels could be as many as millions. Usually the number of images is just hundreds. If we use the signals from all voxels as features for classification, since the number of features is much more than the sample size, it would be a high dimensional problem.

LDA is an asymptotically optimal classifier under traditional large sample scenario, that is, the dimension of variables (p) is fixed and the sample size (n) tends to infinity. However it is not applicable in high dimensional data. Bickel and Levina (2004) [5] demonstrated that LDA performs asymptotically no better than random guessing if $p/n \to \infty$.

As mentioned in 1.3, there are two main issues in high dimensional classification using LDA. The first issue is when $p > n$, the sample covariance $\hat{\Sigma}$ will be singular. To solve this, independence rule (IR) ignores the correlations among features and use diagonal of $\hat{\Sigma}$ to replace $\hat{\Sigma}$, which is applicable for any high dimensions. Bickel and Levina (2004) [5] also shows in theory that IR lead to a better classification result than the naive LDA, where the Moore-Penrose inverse is used to replace $\hat{\Sigma}^{-1}$. Another similar way to solve this issue is nearest shrunken centroid classifier [48]. Fan and Fan (2008) [19] propose feature annealed independence rule (FAIR) that performs feature selection by t-test in addition to IR. However, ignoring the covariance structure of the features, these methods can not asymptotically achieve optimal classification rate.

Many other high dimensional LDA methods are proposed in which the covariance structure in Σ is incorporated. Again, the first challenge for high dimensional LDA is the singularity of $\hat{\Sigma}$. Many methods have been proposed for covariance matrix estimation or precision matrix estimation in high dimension scenario (see, e.g., [6, 7, 8, 10, 11, 12, 44, 45, 47, 53]). The covariance matrix estimated by these methods can be directly used in LDA. However, an accurate estimate of Σ does not necessarily lead to better classification. Fan and Fan (2008)

[19] and Shao *et al.* (2011) [46] shows that even though the true covariance matrix is known, the classification could be no better than random guess because of the noise accumulated from estimating the means. This lead to the second challenge for high dimensional LDA. That is the classification performance is poor due to the noise brought in from the estimate of many non-informative features.

To address this issue, Shao *et al.* [46] assumes sparseness in both $\boldsymbol{\Delta}$, which measures the difference of the two classes in mean and the covariance matrix Σ. Then Σ and $\boldsymbol{\Delta}$ are estimated by hard thresholding for classification. This method is then extended to quadratic discriminant analysis in [33]. One issue with this method is the assumption of the sparseness on the covariance matrix is too restrictive. Witten and Tibshirani (2011) [51] proposed penalized LAD by applying penalties on the discriminant vectors. Cai and Liu (2011) [9], Mai and Zhou (2012) [36] and Fan *et al.* (2012) [20] assume the sparsity of the discriminant direction $\boldsymbol{\beta}^{Bayes} = \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. These methods borrow the idea of penalization in regression to regularize the estimated discriminant direction $\boldsymbol{\beta}^{Bayes}$ directly, avoiding of estimating Σ^{-1}. The same idea is still using in [13]. Though informative direction plays a direct role in the discriminant function, selecting the informative features would be easier for interpretation. Xu *et al.* (2014) [55] proposed a covariance-enhanced method to select informative features for linear discriminant analysis. However the proposed method doesn't directly enforce the selecting of informative features.

In this research, we take advantage of the spatial dependency among features and assume the features are equipped with spatial covariance structure. We then developed a penalized maximum likelihood estimation procedure to simultaneously estimate the covariance parameter, the mean, and select important features for discriminant analysis.

For a spatial domain of interest D in $\mathbb{R}^d$, we consider two classes of spatial processes

$\{y_k(s) : s \in D, k = 1, 2\}$, $(k = 1, 2)$, such that

$$y_k(s) = \mu_k(s) + \epsilon(s), \tag{3.1.1}$$

Where $\mu_k(s)$ is the mean effect function and $\epsilon(s)$ is the corresponding random noise. Assume that the error process $\{\epsilon(s) : s \in D\}$ is a Gaussian process with mean zero and a covariance function

$$\gamma(s, s'; \boldsymbol{\theta}) = cov(\epsilon(s), \epsilon(s')) \tag{3.1.2}$$

where $s, s' \in D$ and $\boldsymbol{\theta}$ is a $q \times 1$ vector of covariance function parameters. We assume the spatial domain is expanding as the number of samples on the domain is increasing:

A 1. Assume the sample set $D \in \mathbb{R}^d$ $(d \geq 1)$ is predetermined and non-random with the restriction $\left\|s_i - s_j\right\|_2 \geq \epsilon > 0$, for $s_i,\ s_j \in D$ for all pairs $i, j = 1, 2, ..., p$ to ensure that the sampling domain increases in extent as p increases.

Assume for any sample of the spatial processes, there are observations at p discrete sites $s_1, ..., s_p \in D$. Suppose $y_{ki}(s)$ $(i = 1, 2, ..., n_1)$ is from class $\mathcal{C}_k$ $(k = 1, 2)$. Let $Y_{kij} = y_{ki}(s_j)$ be the observation at jth site for the ith sample of spatial process $y_k(s)$, where $k = 1, 2$, $i = 1, 2, ..., n_k$, $j = 1, 2, ..., p$, then the jth observation for sample i can be represented by

$$Y_{kij} = \mu_{kj} + \epsilon_{kij} \tag{3.1.3}$$

where $\mu_{kj} = \mu_k(s_j)$ is the mean effect at jth location in class $\mathcal{C}_k$ and $\epsilon_{kij} = \epsilon_{ki}(s_j)$ is the corresponding Gaussian random noise for ith sample at jth location. In matrix notation,

the above model can be written as:

$$\boldsymbol{Y}_{ki} = \boldsymbol{\mu}_k + \boldsymbol{\epsilon}_{ki} \tag{3.1.4}$$

where $\boldsymbol{Y}_{ki} = (Y_{ki1}, ..., Y_{kip})^T$, $\boldsymbol{\mu}_k = (\mu_{k1}, ..., \mu_{kp})^T$ is the mean vector of class $\mathcal{C}_k$ and $\boldsymbol{\epsilon}_{ki} = (\epsilon_{ki1}, ...\epsilon_{kip})^T$ has multivariate nomal distribution $N(\mathbf{0}, \Sigma)$. Since $\epsilon(s)$ has a covariance function (3.1.2), the covariance matrix Σ can be represented by $\Sigma(\boldsymbol{\theta}) = [\gamma(s_i, s_j; \boldsymbol{\theta})]_{i,j=1}^p$, i.e. $\gamma(s_i, s_j)$ is the (i, j)th entry. By (3.1.4), we have

$$\boldsymbol{Y}_{ki} \sim N(\boldsymbol{\mu}_k, \Sigma(\boldsymbol{\theta})) \tag{3.1.5}$$

Assume $\boldsymbol{\theta}_0$ be the true parameter in 3.1.2. If $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we write $\Sigma(\boldsymbol{\theta})$ as Σ for simplicity. We need to make some assumptions on the covariance function $\gamma(s_i, s_j; \boldsymbol{\theta})$:

A 2. (i) Let Ξ be the parameter space for $\boldsymbol{\theta}$. Assume the covariance function $\gamma(s, t; \boldsymbol{\theta})$ is stationary, isotropy, and twice differentiable with respect to $\boldsymbol{\theta}$ for all $\boldsymbol{\theta} \in \Xi$ and $s, t \in D$.
(ii) $\gamma(s, t; \boldsymbol{\theta})$ is positive-definite in the sense that for every finite subset $\{s_1, s_2, ..., s_p\}$ of D the covariance matrix $\Sigma = [\gamma(s_i, s_j; \boldsymbol{\theta})]$ is positive-definite.

Under the stationary and isotropic assumption, we can write $\Sigma(\boldsymbol{\theta}) = [\gamma(h_{ij}; \boldsymbol{\theta})]_{i,j=1}^p$, where $h_{ij} = \left\| s_i - s_j \right\|_2$ is the distance between site s_i and s_j.

Let $\boldsymbol{Y} = (\boldsymbol{Y}_{11}^T, ..., \boldsymbol{Y}_{1n_1}^T, \boldsymbol{Y}_{21}^T, ..., \boldsymbol{Y}_{2n_2}^T)^T$. As defined in (3.1.5), $\boldsymbol{Y}$ has multivariate normal distribution. Then we have the log-likelihood function for $\boldsymbol{\mu}_k$ and $\boldsymbol{\theta}$

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2; \boldsymbol{Y}) = & -\frac{p(n_1 + n_2)}{2} log(2\pi) - \frac{n_1 + n_2}{2} log|\Sigma(\boldsymbol{\theta})| \\ & - \frac{1}{2} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_k)^T \Sigma(\boldsymbol{\theta})^{-1} (\boldsymbol{Y}_{ki} - \boldsymbol{\mu}_k) \end{aligned} \tag{3.1.6}$$

$\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\theta}$ can be estimated by maximum likelihood estimation (MLE) even in high dimensional settings. In the setting of spatial statistical model, the resulting $\Sigma(\hat{\boldsymbol{\theta}})$ is a positive definite matrix. This resolves the first challenge lies in high dimensional classification.

We denote the resulting estimates as $\hat{\boldsymbol{\mu}}_{1MLE}, \hat{\boldsymbol{\mu}}_{2MLE}, \hat{\boldsymbol{\theta}}_{MLE}$, which can be easily plug into 1.3.1 and get the MLE-LDA classifier $\hat{\delta}_{MLE}$:

$$\hat{\delta}_{MLE}(\boldsymbol{X}) = (\boldsymbol{X} - \frac{\hat{\boldsymbol{\mu}}_{1MLE} + \hat{\boldsymbol{\mu}}_{2MLE}}{2})^T \Sigma^{-1}(\hat{\boldsymbol{\theta}}_{MLE})(\hat{\boldsymbol{\mu}}_{1MLE} - \hat{\boldsymbol{\mu}}_{2MLE}).$$

In section 3.2, we investigate the properties of MLE-LDA classifier. We first derived the parameter estimation consistency in MLE for $p/n \to 0$ and $p/n \to C > 0$. Then we show that MLE-LDA is asymptotically optimal if $p/n \to 0$ under some regulation conditions. However, when $p/n \to C > 0$, the MLE-LDA could be no better than random guess even if the true covariance is known unless the signals are very strong. This indicates that feature selection is still necessary for high dimensional case even if we can estimate the covariance parameter consistently.

In section 3.3, we propose to estimate the parameters by penalized MLE (PMLE) by applying a penalty on $\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, which measures the difference of the two classes in mean. We assume the sparseness of $\boldsymbol{\Delta}$, which indicates that only a fraction of the p features are important in classification. Then we derive the parameter estimation consistency and feature selection consistency of PMLE. In the end a classifier (PLDA classifier) is developed using the PMLE and we show that it is asymptotically optimal even if $p/n \to C > 0$.

Simulation study and real data analysis are conducted in section 3.4 and 3.5. All proofs and remarks about assumptions are given in 3.6.

3.2 Classification using maximum likelihood estimation (MLE-LDA)

In this section, we investigate the consistency of parameter estimation in (3.1.6). Also we investigate the classification performance of $\hat{\delta}_{MLE}$. Here are some notations. Let $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12}, ..., \mu_{1p})$, $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, ..., \mu_{2,p})$ and $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, ..., \theta_{0q})$ be the true parameters. Let $\Sigma_k(\boldsymbol{\theta}), k = 1, 2, ..., q$ be the partial derivative of the matrix $\Sigma(\boldsymbol{\theta})$ with respect to θ_k, i.e. $\frac{\partial}{\partial\theta_k}\Sigma(\boldsymbol{\theta}) = \Sigma_k(\boldsymbol{\theta})$. Also let $\Sigma^k(\boldsymbol{\theta}), k = 1, 2, ..., q$ denote the partial derivative of the matrix $\Sigma(\boldsymbol{\theta})^{-1}$ with respect to θ_k, i.e. $\frac{\partial}{\partial\theta_k}\Sigma^{-1}(\boldsymbol{\theta}) = \Sigma^k(\boldsymbol{\theta})$. Also, denote $\Sigma_{kj}(\boldsymbol{\theta}) = \frac{\partial\Sigma(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_j}$ and $\Sigma^{kj}(\boldsymbol{\theta}) = \frac{\partial\Sigma^{-1}(\boldsymbol{\theta})}{\partial\theta_k\partial\theta_j}$. We are going to simplify the notation if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, i.e. we write $\Sigma(\boldsymbol{\theta}_0)$ as Σ, $\Sigma^{-1}(\boldsymbol{\theta}_0)$ as Σ^{-1}, $\Sigma_k(\boldsymbol{\theta}_0)$ as Σ_k and $\Sigma^k(\boldsymbol{\theta}_0)$ as Σ^k. Denote the set of all the eigenvalues of the square matrix A by $\lambda(A)$. Moreover, denote the maximum and minimum eigenvalues of a square matrix A by $\lambda_{max}(A)$ and $\lambda_{min}(A)$, respectively.

Here are the regularity conditions assumed in Theorem 1.

A 3. $\lim sup_{p\to\infty}\lambda_{\max}(\Sigma) < \infty$, $\lim inf_{p\to\infty}\lambda_{\min}(\Sigma) > 0$

A 4. $\|\Sigma_k\|_F^{-2} = O_p(p^{-1})$

A 5. Assume $\lim_{p\to\infty} a_{ij}$ exist, where $a_{ij} = \frac{t_{ij}}{t_{ii}^{1/2}t_{jj}^{1/2}}$ and $t_{ij} = tr(\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_j)$.

A 6. There exists an open subset ω that contains the true parameter point $\boldsymbol{\theta}_0$ such that for all $\boldsymbol{\theta}^* \in \omega$, we have:

(i) $-\infty < \lim_{p\to\infty}\lambda_{\min}(\Sigma_k(\boldsymbol{\theta}^*)) < \lim_{p\to\infty}\lambda_{\max}(\Sigma_k(\boldsymbol{\theta}^*)) < \infty$.

(ii) $-\infty < \lim_{p\to\infty}\lambda_{\min}(\Sigma_{kj}(\boldsymbol{\theta}^*)) < \lim_{p\to\infty}\lambda_{\max}(\Sigma_{kj}(\boldsymbol{\theta}^*)) < \infty$.

(iii) $\| \frac{\partial t_{ij}(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}} \|_2 = O_p(p)$, where $t_{ij}(\boldsymbol{\theta}^*) = tr(\Sigma^{-1}(\boldsymbol{\theta}^*)\Sigma_i(\boldsymbol{\theta}^*)\Sigma^{-1}(\boldsymbol{\theta}^*)\Sigma_j(\boldsymbol{\theta}^*))$

Since $\Sigma^k = -\Sigma\Sigma_k\Sigma$ and $\Sigma^{kj} = \Sigma^{-1}(\Sigma_k\Sigma^{-1}\Sigma_j + \Sigma_j\Sigma^{-1}\Sigma_k - \Sigma_{kj})\Sigma^{-1}$, by A3 and A 6,

we have

$$-\infty < \lim_{p\to\infty} \lambda_{\min}(\Sigma^k(\boldsymbol{\theta}^*)) < \lim_{p\to\infty} \lambda_{\max}(\Sigma^k(\boldsymbol{\theta}^*)) < \infty$$

and

$$-\infty < \lim_{p\to\infty} \lambda_{\min}(\Sigma^{kj}(\boldsymbol{\theta}^*)) < \lim_{p\to\infty} \lambda_{\max}(\Sigma^{kj}(\boldsymbol{\theta}^*)) < \infty.$$

Notice that for any $p \times p$ matrix A we have $\|A\|_F \leq \sqrt{p}\,\|A\|_2 = \sqrt{p}\lambda_{\max}(A)$, then from A 6 we have:

(1) $\left\|\Sigma^k(\boldsymbol{\theta}^*)\right\|_F = O_p(\sqrt{p})$;

(2) $\left\|\Sigma^{kj}(\boldsymbol{\theta}^*)\right\|_F = O_p(\sqrt{p})$.

First we have the following theorem about MLE consistency of (3.1.6).

Theorem 1. Assume A1-A6 hold . Let $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\theta}_0)$ be the true parameter. The maximum likelihood estimation (MLE) of (3.1.6) is: $\hat{\boldsymbol{\mu}}_{1MLE} = \bar{\boldsymbol{Y}}_{1\cdot}$, $\hat{\boldsymbol{\mu}}_{2MLE} = \bar{\boldsymbol{Y}}_{2\cdot}$, $\hat{\boldsymbol{\theta}}_{MLE}$, where $\bar{\boldsymbol{Y}}_{k\cdot} = \sum_{i=1}^{n_k} \boldsymbol{Y}_{ki}/n_k$. Also,

(i) If $p/n \to 0$, $\left\|\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0\right\|_2 = O_p(\frac{1}{\sqrt{np}})$;

(ii) If $p/n \to C$ with $0 < C \leq \infty$ and $\sqrt{p}/n \to 0$, $\left\|\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0\right\|_2 = O_p(\frac{1}{n})$.

Proof. See the proof in the section 3.6. □

Theorem 1 shows that under the spatial statistical model, all the parameters $(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\theta})$ can be estimated consistently by the MLE for either $p/n \to 0$ or p/n goes to a positive constant of ∞. Therefore we obtain a positive definite covariance matrix estimate of $\Sigma(\boldsymbol{\theta})$. We can easily plug the MLE $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\mu}}_{MLE1}$, $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_{MLE2}$ and $\hat{\Sigma} = \Sigma(\hat{\boldsymbol{\theta}}_{MLE})$ into (1.3.7)

to build up the classification function as following:

$$\hat{\delta}_{MLE}(\boldsymbol{X}) = \Big(\boldsymbol{X} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_{1MLE} + \hat{\boldsymbol{\mu}}_{2MLE})\Big)^T \Sigma^{-1}(\hat{\boldsymbol{\theta}}_{MLE})\Big(\hat{\boldsymbol{\mu}}_{1MLE} - \hat{\boldsymbol{\mu}}_{2MLE}\Big) \qquad (3.2.1)$$

Then a new observation $\mathbf{x}$ would be classified into class $\mathcal{C}_1$ if $\hat{\delta}_{MLE}(\mathbf{x}) > 0$ and $\mathcal{C}_2$ otherwise. Using the same notations in section 1.3, the conditional misclassification rate is defined by (1.3.2) and (1.3.4). For simplicity, We are going to use $\hat{\boldsymbol{\mu}}_1$ to denote $\hat{\boldsymbol{\mu}}_{MLE1}$, $\hat{\boldsymbol{\mu}}_2$ to denote $\hat{\boldsymbol{\mu}}_{MLE2}$, $\hat{\boldsymbol{\theta}}$ to denote $\hat{\boldsymbol{\theta}}_{MLE}$ in this section.

We will see in the following theorem that the approximate optimal error rate can be achieved while $p/n \to 0$. However, if $p/n \to C$ with $0 < C \leq \infty$, the error rate would be no better than random guessing even if we know the true covariance, due to the error accumulated in the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

Theorem 2. Let $C_p = \boldsymbol{\Delta}^T \Sigma(\boldsymbol{\theta})\boldsymbol{\Delta}$. Assume $\frac{p}{n} \to 0$, $C_p \to C_0$ with $0 \leq C_0 \leq \infty$, and $nC_p \to \infty$ as $n, p \to \infty$.

(1) The overall misclassification rate $W(\hat{\delta}_{MLE}; \Theta)$ is asymptotically sub-optimal. In other words, $W(\hat{\delta}_{MLE}; \boldsymbol{\theta}) - W_{OPT} \xrightarrow{P} 0$.

(2) Moreover, if $C_p \to C_0$ with $0 \leq C_0 < \infty$ or if $C_p \to \infty$ and $C_p \frac{p}{n} \to 0$, then $W(\hat{\delta}_{MLE}; \boldsymbol{\theta})$ is asymptotically optimal, i.e. $\frac{W(\hat{\delta}_{MLE}; \Theta)}{W_{OPT}} \xrightarrow{P} 1$.

Proof. See the proof in the section 3.6. □

The following theorem shows that while $\frac{p}{n}$ goes to a positive constant or ∞, even though the true covariance is known, the error accumulated in the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ would cause biased misclassification rate unless the signal levels are extremely high. This discovery

suggests that eventhough there's no problem in parameter estimation in our model even in high dimensional case, it is still necessary to select important features for classification.

Theorem 3. Assume the true covariance Σ is known, denote the classifier function as

$$\delta_{\hat{\boldsymbol{\mu}}}(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}})^T \Sigma^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) \tag{3.2.2}$$

where $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$ are MLE in (3.1.6) and $\hat{\boldsymbol{\mu}} = \frac{\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2}{2}$. Assume $p/n \to C$ with $0 < C \leq \infty$, $C_p \to C_0$ with $0 \leq C_0 \leq \infty$. Assume $n_1 \neq n_2$ and $n_k > \frac{n}{4}$ $(k = 1, 2)$, then

(1) For $\frac{C_p}{p/n} \to \infty$, then $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \xrightarrow{P} 0$ and $W_{OPT} \xrightarrow{P} 0$ but $\frac{W(\hat{\delta}_{\hat{\boldsymbol{\mu}}})}{W_{OPT}} \xrightarrow{P} \infty$.

(2) For $\frac{C_p}{p/n} \to c$ with $0 < c < \infty$,

(i) if $\frac{p}{n} \to C < \infty$, then $lim_P W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) > 1 - \Phi(\frac{\sqrt{C_0}}{2})$;

(ii) if $\frac{p}{n} \to \infty$, then $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \xrightarrow{P} 0$ and $W_{OPT} \xrightarrow{P} 0$, but $\frac{W(\hat{\delta}_{\hat{\boldsymbol{\mu}}})}{W_{OPT}} \xrightarrow{P} \infty$.

(3) For $\frac{C_p}{p/n} \to 0$, then $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \xrightarrow{P} \frac{1}{2}$.

Proof. See the proof in the section 3.6. □

Corollary 1. With all conditions the same as in Theorem 3. If $n_1 = n_2$, then

(1) If $\frac{C_p}{\sqrt{p/n}} \to \infty$, then $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \xrightarrow{P} 0$ and $W_{OPT} \xrightarrow{P} 0$, but $\frac{W(\hat{\delta}_{\hat{\boldsymbol{\mu}}})}{W_{OPT}} \xrightarrow{P} \infty$;

(2) If $\frac{C_p}{\sqrt{p/n}} \to c$ with $0 < c < \infty$,

(i) If $\frac{p}{n} \to C$, then $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \xrightarrow{P} 1 - \Phi(\frac{c}{2\sqrt{4 + c/\sqrt{C}}})$ and $W_{OPT} \xrightarrow{P} 1 - \Phi(\frac{\sqrt{c\sqrt{C}}}{2})$

(ii) If $\frac{p}{n} \to \infty$, then $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \xrightarrow{P} 1 - \Phi(\frac{c}{4})$, and $W_{OPT} \xrightarrow{P} 0$;

(3) If $\frac{C_p}{\sqrt{p/n}} \xrightarrow{P} 0$, we have $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \xrightarrow{P} \frac{1}{2}$.

Proof. See the proof in the section 3.6. □

Theorem 3 and Corollary 1 shows that while $p/n \to C$ with $0 < C \leq \infty$, $\hat{\delta}_{\hat{\boldsymbol{\mu}}}$ is never asymptotically optimal. It is asymptotically sub-optimal only if $C_p \to \infty$. It reveals that though there's no difficulty in applying LDA on spatial dependent data if estimate the parameters by MLE, however, in high dimensional case ($p/n \to C$ with $0 < C \leq \infty$), the discriminant performance may be poor due to the noise accumulated in the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ (see, e.g., [19] and [46]). Therefore, feature selection is still critical for classification with features of high dimension. Fan and Fan (2008) [19] seeks to extract salient features by two-sample t-test and proved that t-test can pick up all important features by choosing an appropriate critical value once the features are assumed to be independent. Shao *et al.* (2011) [46] proposes to select features by threshold. For the spatially correlated features, we can use penalized maximum likelihood estimates (PMLE) for feature selection.

3.3 Classification using penalized maximum likelihood estimation (PLDA)

3.3.1 The penalized maximum likelihood estimation (PMLE)

In this section, we consider the high dimensional classification problem (i.e. $p/n \to C$ with $0 < C \leq \infty$ as $p \to \infty$ and $n \to \infty$). We continue to use the notation in section 3.1. $\boldsymbol{\Delta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ is a p dimensional vector. Then $\boldsymbol{\Delta} = (\Delta_1, ..., \Delta_2) = (\mu_{21} - \mu_{11}, ..., \mu_{2p} - \mu_{1p})$ is the difference of the mean effect between class $\mathcal{C}_1$ and $\mathcal{C}_2$. Define the signal set $S = \{j : \Delta_j \neq 0\}$. Let s be the number of non zero elements in $\boldsymbol{\Delta}$. The important features are those in the set S. Instead of assuming the sparsity of discriminant direction as in [9], [20]

and [36], we assume the sparsity of $\boldsymbol{\Delta}$, that is $s \ll n$ and $s/n \to 0$. Then employing the idea of penalization in regression, we develop the penalized maximum likelihood estimates to estimate $\boldsymbol{\Delta}$ and $\boldsymbol{\theta}$. As defined in section 3.1, the observations $\boldsymbol{Y}_{ki}$ are normally distributed $\boldsymbol{Y}_{ki} \sim N(\boldsymbol{\mu}_k, \Sigma(\boldsymbol{\theta}_0))$ for $k = 1, 2$ and $i = 1, 2, ..., n_k$.

Let I_p be the $p \times p$ identity matrix and J_p be the $p \times p$ matrix with 1 as its entries. Define diagonal block matrix for square matrix A as $diag_n(A)$ with $n \times n$ blocks, i.e.

$$diag_n(A) = \underbrace{\begin{pmatrix} A & 0 & \cdots & 0 \\ 0 & A & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & A \end{pmatrix}}_{n \times n \ blocks}$$

Then define $\tilde{I}_{n,p} = diag_n(I_p)$. Also define block matrices $\tilde{J}_{n,p}$ and $\tilde{1}_{n,p}$ as followed:

$$\tilde{J}_{n,p} = \underbrace{\begin{pmatrix} I_p & I_p & \cdots & I_p \\ I_p & I_p & \cdots & I_p \\ \vdots & \vdots & \ddots & \vdots \\ I_p & I_p & I_p & I_p \end{pmatrix}}_{n \times n \ blocks}, \quad \tilde{1}_{n,p} = \underbrace{\begin{pmatrix} I_p \\ I_p \\ \vdots \\ I_p \end{pmatrix}}_{n \ blocks}$$

Let $\boldsymbol{Y} = (\boldsymbol{Y}_{11}^T, \cdots, \boldsymbol{Y}_{1n_1}^T, \boldsymbol{Y}_{21}, \cdots, \boldsymbol{Y}_{2n_2}^T)^T$. In order to estimate $\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, we make transformation with $\boldsymbol{Y}$. Let $\mathbf{Z} = \mathbf{V}\boldsymbol{Y}$, where $\mathbf{V}$ is a $(n-1)p \times np$ matrix made up by the first $(n-1)p$ rows of $\tilde{I}_{n,p} - \frac{1}{n}\tilde{J}_{n,p}$. Then $\mathbf{Z} = (\mathbf{Z}_1^T \ \mathbf{Z}_2^T \cdots \mathbf{Z}_{n-1}^T)^T$, where $\mathbf{Z}_i = \boldsymbol{Y}_{1i} - \bar{\boldsymbol{Y}}$ for $i = 1, 2, ..., n_1$, $\mathbf{Z}_i = \boldsymbol{Y}_{2(i-n_1)} - \bar{\boldsymbol{Y}}$ for $i = n_1+1, n_2+2, ..., n-1$ and $\bar{\boldsymbol{Y}} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k}\boldsymbol{Y}_{ki}$. Then

- $\mathbf{Z}_i \sim N(-\tau_2\boldsymbol{\Delta}, \frac{n-1}{n}\Sigma(\boldsymbol{\theta}_0))$ for $i = 1, 2, ..., n_1$;

- $\mathbf{Z}_i \sim N(\tau_1\boldsymbol{\Delta}, \frac{n-1}{n}\Sigma(\boldsymbol{\theta}_0))$ for $i = n_1 + 1, ..., n - 1$.

where $\tau_1 = \frac{n_1}{n}$ and $\tau_2 = \frac{n_2}{n}$. Also, $cov(\mathbf{Z}_i, \mathbf{Z}_j) = -\frac{1}{n}\Sigma$ for $i \neq j$. Define $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ such that $\boldsymbol{X}_i = \boldsymbol{X}^{(1)}$ for $i = 1, 2, ..., n_1$ and $\boldsymbol{X}_i = \boldsymbol{X}^{(2)}$ for $i = n_1 + 1, , ..., n - 1$, where

$$\boldsymbol{X}^{(1)} = \begin{pmatrix} -\tau_2 & 0 & \cdots & 0 \\ 0 & -\tau_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & -\tau_2 \end{pmatrix}_{p\times p}, \quad \boldsymbol{X}^{(2)} = \begin{pmatrix} \tau_1 & 0 & \cdots & 0 \\ 0 & \tau_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \tau_1 \end{pmatrix}_{p\times p}$$

Write $\boldsymbol{X} = (\boldsymbol{X}_1^T \boldsymbol{X}_2^T \cdots \boldsymbol{X}_{n-1}^T)^T$ and write $\boldsymbol{\beta} = \boldsymbol{\Delta}$. Then $\mathbf{Z}$ is a $(n-1)p \times 1$ vector with multivariate normal distribution $N(\boldsymbol{X}\boldsymbol{\beta}, \dot{\Sigma})$, where $\dot{\Sigma} = (\tilde{I}_{n-1,p} - \frac{1}{n}\tilde{J}_{n-1,p})diag_{n-1}(\Sigma)$. Denote all the unknown parameters by $\boldsymbol{\eta} = (\boldsymbol{\beta}, \boldsymbol{\theta}) \in \mathbb{R}^{p+q}$. Since

$$\left|\dot{\Sigma}\right| = \left|\tilde{I}_{n-1,p} - \frac{1}{n}\tilde{J}_{n-1,p}\right| |diag_{n-1}\Sigma(\boldsymbol{\theta}))| = (\frac{1}{n})^p \left|\Sigma(\boldsymbol{\theta})\right|^{n-1}$$

and $(\tilde{I}_{n-1,p} - \frac{1}{n}\tilde{J}_{n-1,p})^{-1} = \tilde{I}_{n-1,p} + \tilde{J}_{n-1,p}$, we can write the penalized log-likelihood function for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{Z}) = -\frac{np}{2}log(2\pi) - \tfrac{1}{2}log\left|\dot{\Sigma}\right| - \tfrac{1}{2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^T\dot{\Sigma}^{-1}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}) - n\sum_{j=1}^{p} P_\lambda(|\beta_j|) \tag{3.3.1}$$

$$\begin{aligned} =& C_{n,p} - \frac{n-1}{2}log\left|\Sigma\right| - \tfrac{1}{2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^T diag_{n-1}(\Sigma^{-1})(\tilde{I}_{n-1,p} + \tilde{J}_{n-1,p})(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}) \\ & - n\sum_{j=1}^{p} P_\lambda(|\beta_j|) \end{aligned}$$

where $C_{n,p} = -\frac{(n-1)p}{2}\log\pi + \frac{p}{2}\log n$. $P_\lambda(x)$ is a generic sparsity-inducing penalty. It could be the lasso penalization or folded concave penalization (such as the SCAD and the MCP). We use SCAD penalization in this research.

The true parameter $\boldsymbol{\beta}_0$ is a column vector parameters of size p, and $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, ..., \theta_{0q})$ is the $q-$dimensional parameters in covariance function. Assume $\boldsymbol{\beta}_0$ is sparse. Without loss of generality we can write $\boldsymbol{\beta}_0 = (\beta_{0,1}^T, \beta_{0,2}^T)^T$, where $\beta_{0,1} \in \mathbb{R}^s$ is non-zero component, $\beta_{0,2} = \mathbf{0}_{(p-s)\times 1}$ is the zero component of β_0 and $\frac{s}{n} \to 0$ as $n, p, s \to \infty$. Also, we can write $\boldsymbol{X}_i = (\boldsymbol{X}_i^1, \boldsymbol{X}_i^2)$, for $i = 1, 2, .., n$, where $\boldsymbol{X}_i^1$ is the $p \times s$ submatrix of $\boldsymbol{X}_i$ formed by columns in $supp(\boldsymbol{\beta}_0)$ and $\boldsymbol{X}_i^2$ is the $p \times (p-s)$ complement matrix.

We are going to follow the one step estimation procedure in [14]. As demonstrated in [58], the one-step method is as efficient as the fully interative method both empirically and theoretically, provided that the initial estimators are reasonably good. Here's the algorithm:

Algorithm

1. Initialize $\boldsymbol{\beta}$ by minimizing $R(\boldsymbol{\beta}) = (\mathbf{Z} - \boldsymbol{X}\beta)^T(\mathbf{Z} - \boldsymbol{X}\beta) + n\sum_{j=1}^p P_\lambda(|\beta_j|)$ with respect to $\boldsymbol{\beta}$. Denote the result by $\hat{\boldsymbol{\beta}}^{(0)}$;

2. Fix $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(0)}$, estimate $\boldsymbol{\theta}$ by maximize $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(0)}; \mathbf{Z})$ in (3.3.1) with respect to $\boldsymbol{\theta}$. Denote the result by $\hat{\boldsymbol{\theta}}^{(0)}$;

3. Fix $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(0)}$, update $\boldsymbol{\beta}$ by maximize $Q(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\beta}; \mathbf{Z})$ in (3.3.1) with respect to $\boldsymbol{\beta}$. Denoted the result by $\hat{\boldsymbol{\beta}}^{(1)}$;

4. Fix $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(1)}$, estimate $\boldsymbol{\theta}$ by maximize $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(1)}; \mathbf{Z})$ in (3.3.1) with respect to $\boldsymbol{\theta}$. Denote the result by $\hat{\boldsymbol{\theta}}^{(1)}$.

Then $\hat{\boldsymbol{\theta}}_{PMLE} = \hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}_{PMLE} = \hat{\boldsymbol{\beta}}^{(1)}$ are the desired estimates. We call $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(1)}$

one-step PMLE.

Denote $\hat{\boldsymbol{\Delta}}_{PMLE} = \hat{\boldsymbol{\beta}}^{(1)}$ and $\boldsymbol{\theta}_{PMLE} = \hat{\boldsymbol{\theta}}^{(1)}$ from the algorithm. Then $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ can be estimated by $\hat{\boldsymbol{\mu}}_{1PMLE} = \bar{\boldsymbol{Y}} - \tau_2\hat{\boldsymbol{\Delta}}_{PMLE}$ and $\hat{\boldsymbol{\mu}}_{2PMLE} = \bar{\boldsymbol{Y}} + \tau_1\hat{\boldsymbol{\Delta}}_{PMLE}$. Also we have estimated covariance $\hat{\Sigma} = \Sigma(\hat{\boldsymbol{\theta}}_{PMLE})$, where the (i,j)th element of $\hat{\Sigma}$ is:

$$\hat{\sigma}_{i,j} = \gamma(h_{ij}; \hat{\boldsymbol{\theta}}_{PMLE}) \tag{3.3.2}$$

where $h_{ij} = \left\|s_j - s_i\right\|_2$ is the distance between site s_i and s_j.

3.3.1.1 Consistency of PMLE

Penality function largely determines the sampling properties of the penalized likelihood estimators. Some additional assumptions about the penality function and tuning parameter λ are needed:

A 7. Assume $a_n = O_p(\frac{1}{\sqrt{n}})$, where $a_n = \max_{1\le j\le p}\{p'_{\lambda_n}(|\beta_{0j}|), \beta_{0j} \neq 0\}$

A 8. $b_n \to 0$ as $n \to \infty$, where $b_n = \max_{1\le j\le m}\{p''_{\lambda_n}(|\beta_{0j}|), \beta_{0j} \neq 0\}$

A 9. $\lambda_n \to 0$ and $\lambda_n/\sqrt{\frac{s}{n}} \to \infty$.

A 10. $\liminf_{\substack{n\to\infty \\ p\to\infty}} \liminf_{\theta\to 0+} P'_{\lambda_n}(|\theta|)/\lambda_n > 0$

A 7 ensure the unbiasedness property for large parameters and the existence of the consistent penalized likelihood estimator. A 8 ensure that the penalty function does not influence the penalized likelihood estimators more than the likelihood function. A10 makes the penalized likelihood estimators possess the sparsity property and A9 lead to the variable selection consistency.

Smoothly Clipped Absolute Deviation (SCAD) penality satisfy all those assumptions and we are going to apply SCAD in the simulation and real data analysis in this paper. Fan

and Li (2001) [21] proposed the SCAD penality function and claims that it satisfy three good properties: unbiasedness, sparsity and continuity. Unbiasedness means there is no overpenalization of large features to avoid unnecessary modeling biases. Sparsity means the insignificant parameters are set to 0 by a thresholding rule to reduce model complexity. Continuity makes sure the penalized likelihood produces continuous estimators. It is defined as

$$p_\lambda(\beta) = \begin{cases} \lambda\,|\beta| & \text{if } |\beta| \le \lambda \\ -\frac{\beta^2 - 2a\lambda\beta + \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta| \le a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

for some $a > 0$. More details can be found in [21].

Theorem 4. Assume conditions A 1-A 10 hold. Assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{1,0}^T, \boldsymbol{\beta}_{2,0}^T)^T$, where $\boldsymbol{\beta}_{1,0} \in \mathbb{R}^s$ is non-zero component, $\boldsymbol{\beta}_{2,0} = \mathbf{0}_{(p-s)\times 1}$ is the zero component of $\boldsymbol{\beta}_0$ with $\frac{s}{n} \to 0$, $\frac{p}{n} \to C$ with $0 < C \le \infty$ as $n, p, s \to \infty$. The PMLE estimate of (3.3.1) from the Algorithm in section 3.3.1 is $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ with $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ and $\hat{\boldsymbol{\beta}}_1$ is a subvector of $\hat{\boldsymbol{\beta}}$ formed by components in $supp(\boldsymbol{\beta}_0)$. Then $\hat{\boldsymbol{\eta}}$ satisfy:

a. **(consistancy)** $\left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_2 = O_p(\frac{1}{\sqrt{np}})$ and $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right\|_2 = O_p(\sqrt{\frac{s}{n}})$.

b. **(sparsity)** $\hat{\boldsymbol{\beta}}_2 = 0$ with probability tending to 1 as $n \to \infty$.

Proof. See the proof in the section 3.6. □

3.3.1.2 Covariance tappering and PMLE

When the number of sites is large (p is large) for each realization of the spatial process, calculating the likelihood can be computationally infeasible (requiring $\mathcal{O}(p^3)$ calculation). Covariance tapering can be used to approximate the likelihood. While the covariance matrix is replaced by a tapered one, the resulting matrixes can then be manipulated using efficient sparse matrix algorithms which would reduce computation effectively.

In section 3.1, the covariance matrix is defined as $\Sigma(\boldsymbol{\theta}) = [\gamma(s_i, s_j; \boldsymbol{\theta})]_{i,j=1}^{p}$. Under A2, we can simply write it as $\Sigma = [\gamma(h_{ij}; \boldsymbol{\theta})]_{i,j=1}^{p}$, where $h_{ij} = \left\| s_i - s_j \right\|_2$ is the distance between sites s_i and s_j. Let $K_T(h, w)$ denote a tapering function, which is an isotropic autocorrelation function when $0 < d < w$ and 0 when $h \geq w$ for a given threshold $w > 0$. Compactly correlation function can be used as the tapering function. We are going to use an tapering function from [50],

$$K_T(h, \omega) = [(1 - h/w)_+]^2 \tag{3.3.3}$$

where $x_+ = \max(x, 0)$ in which case the correlation is 0 at lag distance greater than the threshold distance w. Let $\mathbf{K}(w) = [K_T(h_{ii'}, w)]_{i,i'=1}^{p}$ denote the $p \times p$ tapering matrix. Then a tapered covariance of Σ is defined as $\Sigma_T = \Sigma \circ \mathbf{K}(w)$, where $\circ$ is the Schur product (i.e. elementwise product). By the properties of the Schur product (see, e.g., [26], chap. 5), the tapered covariance matrix would keep the positive definiteness thus it is still a valid covariance matrix. When p is large, we are going to approximate the penalized log-likelihood

(3.3.1) by replacing Σ by Σ_T and obtain a covariance tapered penalized logliklihood:

$$Q_T(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{Z}) = -\frac{np}{2} log(2\pi) - \tfrac{1}{2} log \left|\dot{\Sigma}_T\right| - \tfrac{1}{2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^T \dot{\Sigma}_T^{-1}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}) - n\sum_{j=1}^{p} P_\lambda(|\beta_j|) \tag{3.3.4}$$

$$\begin{aligned} =& C_{n,p} - \frac{n-1}{2} log\,|\Sigma_T| - \tfrac{1}{2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^T diag_{n-1}(\Sigma_T^{-1})(\tilde{I}_{n-1,p} + \tilde{J}_{n-1,p})(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}) \\ & - n\sum_{j=1}^{p} P_\lambda(|\beta_j|) \end{aligned}$$

where $C_{n,p} = -\frac{(n-1)p}{2}\log\pi + \frac{p}{2}\log n$.

We keep all the notations the same as in (3.3.1), except Σ is replaced by Σ_T. Let $\boldsymbol{\Delta}_{PMLE,T} = \boldsymbol{\beta}_{PMLE,T}$ and $\boldsymbol{\theta}_{PMLE,T}$ be the penalized maximum likelihood estimates with tapered covariance (PMLE$_T$). We are going to check the consistency of PMLE$_T$. Let $\gamma_k(\boldsymbol{\theta}, h) = \frac{\partial\gamma(\boldsymbol{\theta},h)}{\partial\theta_k}(\boldsymbol{\theta})$ and $\gamma_{jk}(\boldsymbol{\theta}, h) = \frac{\partial^2\gamma(\boldsymbol{\theta},h)}{\partial\theta_k\theta_j}$. Two additional assumptions are made here for regularity.

A 11. Assume $0 < \inf_p\{\frac{w_p}{p^{\delta_0}}\} < \sup_p\{\frac{w_p}{p^{\delta_0}}\} < \infty$, where w_p is the threshold distance in the tapering function for $\delta_0 > 0$.

A 12. Let d ($d \geq 1$) be the dimension of the domain, i.e. $D \subset \mathbb{R}^d$. Assume for all $\boldsymbol{\theta} \in \Xi$ and $1 \leq k, j \leq q$, we have $\gamma(\boldsymbol{\theta}, h), \gamma_k(\boldsymbol{\theta}, h), \gamma_{jk}(\boldsymbol{\theta}, h)$ belong to the function space $\mathcal{L}$, where $\mathcal{L} = \{f(h) : \int_0^\infty h^d f(h)dh < \infty\}$.

Let Σ be the covariance matrix and Σ_T be the tapered covariance matrix. $\Sigma_{k,T} = \frac{\partial\Sigma_T}{\partial\theta_k}$ and $\Sigma_{jk,T} = \frac{\partial^2\Sigma_T}{\partial\theta_j\partial\theta_k}$. By using the tapering function (3.3.3), we have the following result.

Theorem 5. Assume conditions 1-12 hold. Assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{1,0}^T, \boldsymbol{\beta}_{2,0}^T)^T$, where $\boldsymbol{\beta}_{1,0} \in \mathbb{R}^s$ is non-zero component, $\boldsymbol{\beta}_{2,0} = \mathbf{0}_{(p-s)\times 1}$ is the zero component of $\boldsymbol{\beta}_0$ with $\frac{s}{n} \to 0$, $\frac{p}{n} \to C$ with $0 < C \leq \infty$ as $n, p, s \to \infty$. The PMLE estimates of (3.3.4) from Algorithm in section

3.3.1 is $\hat{\boldsymbol{\eta}}_T = (\hat{\boldsymbol{\beta}}_T, \hat{\boldsymbol{\theta}}_T)$ with $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{1,T}^T, \hat{\boldsymbol{\beta}}_{2,T}^T)^T$ and $\hat{\boldsymbol{\beta}}_{1,T}$ is a subvector of $\hat{\boldsymbol{\beta}}_T$ formed by components in $supp(\boldsymbol{\beta}_0)$. Then $\hat{\boldsymbol{\eta}}_T$ satisfy:

a. **(consistancy)** $\left\|\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0\right\|_2 = O_p(\frac{1}{\sqrt{np}})$ and $\left\|\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_0\right\|_2 = O_P(\sqrt{\frac{s}{n}})$.

b. **(sparsity)** $\hat{\boldsymbol{\beta}}_{2,T} = 0$ with probability tending to 1 as $n \to \infty$.

Proof. See the proof in the section 3.6. □

3.3.2 The penalized maximum likelihood estimation LDA (PLDA) classifier

No matter using PMLE or PMLE$_T$, we have the estimates for $\boldsymbol{\Delta}$ and $\boldsymbol{\theta}$ denoted by $\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{\theta}}$. the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are $\hat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{Y}} - \tau_2\hat{\boldsymbol{\Delta}}$ and $\hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{Y}} + \tau_1\hat{\boldsymbol{\Delta}}$. Also we have estimated covariance $\hat{\Sigma} = \Sigma(\hat{\boldsymbol{\theta}})$, where the (i,j)th element of $\hat{\Sigma}$ is:

$$\hat{\sigma}_{ij} = \gamma(\left|s_j - s_i\right|; \hat{\boldsymbol{\theta}}) \tag{3.3.5}$$

Since $p > n$, the error accumulated in estimate of each $\hat{\sigma}_{ij}$ may also cause problems in classification (see, e.g., [6] and [46]). For regularization of the covariance matrix, we use the tapered coavariance matrix in classification function. Specifically, we define $\tilde{\Sigma} = \Sigma_T(\hat{\boldsymbol{\theta}}) = \Sigma(\hat{\boldsymbol{\theta}}) \circ \mathbf{K}(w)$, where $\mathbf{K}(w)$ is defined in section 3.3.1.2. We then replace $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma$ in LDA (1.3.7) by $\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2$ and $\tilde{\Sigma}$ for classification. Then the PLDA function is:

$$\hat{\delta}_{PLDA}(\boldsymbol{X}) = (\boldsymbol{X} - \bar{\boldsymbol{Y}} - \frac{n_1 - n_2}{2n}\hat{\boldsymbol{\Delta}})^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} \tag{3.3.6}$$

where $\bar{\boldsymbol{Y}} = \frac{1}{n}\sum_{k=1}^{2}\sum_{i=1}^{n_k}\boldsymbol{Y}_{ki}$.

The conditional misclassification rate for class 1 and class 2 are defined by (1.3.2) and (1.3.4) with $\hat{\Sigma}$ replaced by $\tilde{\Sigma}$. Similarly we have the overall misclassification rate defined in (1.3.6).

A 13. Assume $\int_1^\infty h^d r(h;\boldsymbol{\theta})dh < \infty$ and $\int_0^1 h^{d-1} r(h;\boldsymbol{\theta})dh < \infty$ for $\boldsymbol{\theta} \in \Xi$.

A 14. Assume there exist a constant M such that for any $h \geq 0$ and $\boldsymbol{\theta} \in \Xi$, $\| \frac{\partial \gamma(h;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \|_2 \leq M$.

Theorem 6. Assume $\hat{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\Delta}}$ in (3.3.6) are estimated from Theorem 4 and 5. Suppose assumptions A1-A3 and A13-14 hold. Assume $\frac{s}{n} \to 0$, $\frac{p}{n} \to C$ with $0 < C \leq \infty$, $C_p \to C_0$ with $0 \leq C_0 \leq \infty$, $\frac{C_p}{\sqrt{s/n}} \to 0$. Also, assume $w = O((\sqrt{np})^{\frac{\alpha}{d}})$ with $0 < \alpha < 1$, and $w^{-1} = O(p^{-\delta})$ with $\delta > 0$, where d is the dimension of the domain. Then the classification error rate of $\hat{\delta}_{PLDA}$ is asymptotically sub-optimal, i.e. $W(\hat{\delta};\Theta) \xrightarrow{P} 1 - \Phi(\frac{\sqrt{C_0}}{2})$. Moreover,

(1) If $C_p \to C_0 < \infty$, $W(\hat{\delta};\Theta)$ is asymptotically optimal, i.e. $\frac{W(\hat{\delta};\Theta)}{W_{OPT}} \xrightarrow{P} 1$;

(2) If $C_p \to \infty$ and $C_p \kappa_{n,p} \to 0$, $W(\hat{\delta};\Theta)$ is asymptotically optimal, i.e. $\frac{W(\hat{\delta};\Theta)}{W_{OPT}} \xrightarrow{P} 1$, where $\kappa_{n,p} = \max(\frac{w^d}{\sqrt{np}}, \frac{1}{w}, \sqrt{\frac{s}{n}})$.

Proof. See the proof in the section 3.6. □

3.4 Simulation Analysis

We do the simulation analysis in this section. Assume the spatial domain of interest D in $\mathbb{R}^2$ is a $u \times u$ square. We can observe signal at each lattice. Then we have $p = u \times u$ features for classification. Assume the mean effects of the signal for class $\mathcal{C}_1$ and $\mathcal{C}_2$ are $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. We assume $\boldsymbol{\mu}_1 = (\mathbf{1}_{10}, \mathbf{0}_{p-10})$ and $\boldsymbol{\mu}_2 = \mathbf{0}_p$, where where $\mathbf{1}_k$ is a k dimension vector with all the elements equal to 1 and $\mathbf{0}_k$ is a k dimension vector with all the elements equal to 0.

4	8	12	16
3	7	11	15
2	6	10	14
1	5	9	13

1	1	0	0
1	1	0	0
1	1	1	0
1	1	1	0

0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0

Figure 3.1 Two dimensional domain example. Left: 2D domain with p=4 $\times$ 4; middle: $\boldsymbol{\mu}_1$; right: $\boldsymbol{\mu}_2$.

For example, if $u = 4$ hence $p = 16$, then the corresponding spatial domain D, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are as in Figure 3.1. In the simulation setting, we let $p = 36, 400, 1225$ respectively.

For the spatial covariance, we generate the error terms from stationary and isotropic Gaussian process with zero-mean. *Matérn* covariance function which was defined in 1.2.1 is widely used as spatial covariance function. We consider two special cases of *Matérn* covariance function: exponential covariance function and polynomial covariance function. Let h be the Euclid distance between two sites on the domain D. Specifically, on the domain $D \in \mathbb{R}^2$, the distance between site i with coordinate $s_i = (x_i, y_i)$ and site j with coordinate $s_j = (x_j, y_j)$ is $h_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.

- **Exponential covariance function** If let the smoothness parameter $\nu = \frac{1}{2}$ in 1.2.1, the the spatial dependence of the error terms reduced to an exponential covariance function:

$$\gamma(h) = \begin{cases} \sigma^2(1-c)exp(-h/r) & \text{if } h > 0 \\ \sigma^2 & \text{if } h = 0 \end{cases}$$

  where σ^2 is the variance parameter, c is the nugget effect and r is the range parameter. In the simulation, we set $\sigma^2 = 1$, $c = 0.2$ and $r = 1, 2, ..., 8, 9$. The classification performance are in Table 3.1, 3.2 and 3.3. The parameter estimation results are in

Table 3.4. The model selection results are in Table 3.5, 3.6, 3.7.

- **Polynomial covariance function** The covariance function of polynomial covariance function is $\gamma(h) = \sigma^2\rho^h$, where ρ is the correlation parameter. Then the covariance matrix is:

$$\Sigma_{p\times p} = \sigma^2 \begin{bmatrix} 1 & \rho^{h_{12}} & \dots & \rho^{h_{1p}} \\ \rho^{h_{21}} & 1 & \dots & \rho^{h_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{h_{p1}} & \rho^{h_{p2}} & \dots & 1 \end{bmatrix}$$

  In the simulation, we let $\sigma^2 = 1$ and $\rho = 0, 0.1, 0.2..., 0.8, 0.9$. Polynomial correlation function ρ^h is a special case of exponential correlation function $exp(-h/r)$, since $\rho^h = exp(-h\log\frac{1}{\rho})$. The classification performance are in Table 3.8, 3.9 and 3.10. The parameter estimation results are in Table 3.11. The model selection results are in Table 3.12, 3.13, 3.14.

100 groups of training sets with $n_1 = n_2 = 30$ are generated according to different setting of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\Sigma(\boldsymbol{\theta}_0)$. For each training set, we can estimate the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\theta}_0$ by MLE, tapered MLE, PMLE and tapered PMLE. 100 groups of testing data sets with $n_1 = n_2 = 100$ are generated for testing the classification performance. The average classification error rate was calculated from the 100 groups of testing data sets. All the reported numbers in the tables are means with their standard errors, calculated by 100 groups of training and testing data sets, in parentheses.

We name the classification method proposed in the research as PLDA. For each choice of p, we compared the classification performance of PLDA with oracle classification, MLE-LDA, PREG-LDA, FAIR (Feature Annealed Independence Rule; [19]) and NB (Naive Bayes;

[5]).

Specifically, MLE-LDA uses $\hat{\boldsymbol{\mu}}_{1MLE}$, $\hat{\boldsymbol{\mu}}_{2MLE}$ and $\Sigma(\hat{\boldsymbol{\theta}}_{MLE})$ in LDA function for classification; PREG-LDA uses $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^{(0)}$ in LDA in classification function, where $\hat{\boldsymbol{\Delta}}^{(0)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ are the same $\hat{\boldsymbol{\Delta}}^{(0)}$ and $\hat{\boldsymbol{\theta}}^{(0)}$ estimated in the algorithm in section 3.3.1. This method doesn't consider spatial correlation in feature selection. NB uses sample mean $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$ and diagnol of sample covariance $\hat{\Sigma}$ in LDA. This method is also known as independent rule(IR). FAIR assumes independence between variables and utilizes t-test for variable selection in IR. Oracle uses true mean $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and true covariance $\Sigma(\boldsymbol{\theta}_0)$. We also compared the average number of variables selected from PMLE, PREG and FAIR. The tunning parameter λ in $PREG$ and $PMLE$ is selected by 10 fold cross validation by minimizing the classification error rate.

For $p = 36$, the performance of MLE-LDA, PREG-LDA, PMLE, FAIR, NB are shown. For $p = 400$ the performance of MLE-LDA, PREG-LDA, PMLE, FAIR, NB are shown. The classification performance of MLE-LDA, PLDA with the parameters estimated by tapered MLE and tapered PMLE are also shown for $p = 400$. For $p = 1225$, the classification performance of the tapered MLE-LDA, tapered PLDA, PREG-LDA, FAIR and NB are shown.

Table 3.1 Classification Accuracy Rate (Exponential covariance, p=36)

	TURE	MLE	PREG	PMLE	FAIR	NB
	p=36					
r=1	0.88(0.02)	0.84(0.03)	0.84(0.04)	0.84(0.05)	0.81(0.04)	0.84(0.04)
r=2	0.88(0.02)	0.84(0.03)	0.84(0.04)	0.84(0.05)	0.76(0.05)	0.78(0.05)
r=3	0.89(0.02)	0.85(0.03)	0.85(0.04)	0.85(0.05)	0.74(0.04)	0.77(0.04)
r=4	0.90(0.02)	0.87(0.03)	0.87(0.03)	0.87(0.04)	0.73(0.05)	0.76(0.05)
r=5	0.91(0.02)	0.88(0.02)	0.88(0.02)	0.88(0.04)	0.72(0.05)	0.75(0.05)
r=6	0.92(0.02)	0.89(0.02)	0.89(0.02)	0.89(0.04)	0.72(0.06)	0.75(0.06)
r=7	0.93(0.02)	0.90(0.02)	0.90(0.03)	0.90(0.03)	0.71(0.06)	0.75(0.06)
r=8	0.93(0.02)	0.91(0.02)	0.91(0.02)	0.91(0.03)	0.71(0.06)	0.75(0.06)
r=9	0.94(0.02)	0.91(0.02)	0.92(0.02)	0.92(0.02)	0.71(0.05)	0.75(0.05)

Table 3.2 Classification Accuracy Rate (Exponential covariance, p=400)

	TURE	MLE	MLE_T	PREG	$PREG_T$	PMLE	$PMLE_T$	FAIR	NB
	p=400								
r=1	0.92(0.02)	0.74(0.03)	0.75(0.03)	0.84(0.05)	0.84(0.05)	0.82(0.05)	0.84(0.04)	0.83(0.04)	0.74(0.04)
r=2	0.92(0.02)	0.75(0.03)	0.76(0.03)	0.85(0.05)	0.85(0.05)	0.85(0.05)	0.86(0.05)	0.79(0.04)	0.67(0.04)
r=3	0.94(0.02)	0.78(0.03)	0.78(0.03)	0.86(0.05)	0.87(0.05)	0.88(0.04)	0.89(0.05)	0.77(0.04)	0.63(0.05)
r=4	0.95(0.01)	0.80(0.03)	0.80(0.03)	0.88(0.04)	0.89(0.05)	0.91(0.03)	0.91(0.03)	0.75(0.04)	0.61(0.05)
r=5	0.95(0.01)	0.81(0.03)	0.82(0.03)	0.90(0.05)	0.90(0.04)	0.92(0.03)	0.93(0.03)	0.74(0.04)	0.60(0.05)
r=6	0.96(0.01)	0.83(0.03)	0.84(0.03)	0.91(0.05)	0.91(0.04)	0.93(0.02)	0.94(0.02)	0.73(0.05)	0.59(0.05)
r=7	0.96(0.01)	0.84(0.03)	0.85(0.02)	0.91(0.04)	0.92(0.04)	0.95(0.02)	0.94(0.02)	0.72(0.06)	0.58(0.05)
r=8	0.97(0.01)	0.85(0.03)	0.86(0.02)	0.92(0.04)	0.93(0.04)	0.95(0.02)	0.95(0.02)	0.72(0.06)	0.58(0.05)
r=9	0.97(0.01)	0.86(0.02)	0.87(0.02)	0.93(0.04)	0.93(0.04)	0.96(0.02)	0.95(0.02)	0.72(0.05)	0.58(0.05)

Finally, the Table 3.15 and Table 3.16 and Table 3.17 showed the classification performance when the covariance are misspecified. We generate data using Gaussian covariance function with $\sigma^2 = 1,\ c = 0.2$, and $r = 1, 2, ..., 9$. Then we estimate the parameters using exponential covariance function. It shows the PLDA classification method still yielded the best performance in the misspecified case.

Table 3.3 Classification Accuracy Rate (Exponential covariance, p=1225)

	TURE	MLE_T	$PREG_T$	$PMLE_T$	FAIR	NB
	p=1225					
r=1	0.92(0.02)	0.67(0.03)	0.83(0.05)	0.80(0.06)	0.83(0.05)	0.66(0.05)
r=2	0.92(0.02)	0.68(0.03)	0.83(0.06)	0.82(0.06)	0.78(0.05)	0.61(0.05)
r=3	0.93(0.02)	0.70(0.04)	0.85(0.06)	0.86(0.05)	0.76(0.04)	0.58(0.04)
r=4	0.94(0.02)	0.72(0.04)	0.87(0.06)	0.89(0.03)	0.75(0.05)	0.57(0.05)
r=5	0.95(0.02)	0.73(0.04)	0.88(0.05)	0.91(0.03)	0.74(0.05)	0.56(0.05)
r=6	0.96(0.01)	0.75(0.03)	0.89(0.06)	0.92(0.03)	0.73(0.04)	0.55(0.04)
r=7	0.96(0.01)	0.76(0.04)	0.89(0.06)	0.93(0.02)	0.73(0.04)	0.55(0.04)
r=8	0.96(0.01)	0.77(0.04)	0.90(0.05)	0.93(0.02)	0.72(0.05)	0.54(0.05)
r=9	0.97(0.01)	0.77(0.04)	0.91(0.05)	0.94(0.02)	0.72(0.05)	0.54(0.05)

Table 3.4 Parameter estimation (Exponential covariance)

			p=36		p=400				p=1225	
		TURE	MLE	PMLE	MLE	PMLE	MLE_T	$PMLE_T$	MLE_T	$PMLE_T$
r=1	r	1	1.0(0.2)	1.0(0.2)	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.2(0.1)	1.2(0.1)
	c	0.2	0.2(0.1)	0.2(0.1)	0.2(0.0)	0.2(0.0)	0.3(0.1)	0.3(0.0)	0.5(0.0)	0.5(0.0)
	σ	1	1.0(0.0)	1.0(0.0)	1.0(0.0)	1(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)
r=2	r	2	2.0(0.3)	2.1(0.3)	2(0.1)	2.0(0.1)	2.1(0.1)	2.1(0.1)	2.5(0.1)	2.5(0.1)
	c	0.2	0.2(0.1)	0.2(0.1)	0.2(0.0)	0.2(0.0)	0.3(0.0)	0.3(0.0)	0.5(0.0)	0.5(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)
r=3	r	3	3.0(0.4)	3.1(0.4)	3.0(0.1)	3.0(0.1)	3.2(0.2)	3.2(0.2)	3.9(0.2)	3.9(0.2)
	c	0.2	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.3(0.0)	0.3(0.0)	0.5(0.0)	0.5(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)
r=4	r	4	4.1(0.6)	4.1(0.6)	4.0(0.2)	4.0(0.2)	4.6(0.4)	4.6(0.4)	5.6(0.4)	5.6(0.4)
	c	0.2	0.2(0.0)	0.2(0.0)	0.2(0.01)	0.2(0.01)	0.3(0.0)	0.3(0.0)	0.5(0.0)	0.5(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)
r=5	r	5	5.1(0.8)	5.1(0.8)	5.0(0.3)	5.0(0.3)	6.4(0.8)	6.4(0.8)	7.6(0.7)	7.6(0.7)
	c	0.2	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.3(0.0)	0.2(0.0)	0.5(0.0)	0.5(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	0.9(0.0)	1.0(0.0)
r=6	r	6	6.1(0.9)	6.1(1.0)	6.1(0.4)	6.1(0.4)	8.8(1.4)	8.8(1.4)	10.1(1.2)	10.1(1.1)
	c	0.2	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.5(0.0)	0.5(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	0.9(0.0)	1.0(0.0)
r=7	r	7	7.1(1.2)	7.1(1.2)	7.1(0.5)	7.1(0.5)	12.0(2.5)	12.0(2.6)	13.3(1.9)	13.3(1.9)
	c	0.2	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.5(0.0)	0.4(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.1)	1.0(0.1)	1.0(0.1)	0.9(0.0)	1.0(0.0)
r=8	r	8	8.1(1.4)	8.1(1.4)	8.1(0.6)	8.1(0.6)	16.6(4.6)	16.8(4.9)	17.6(3.3)	17.7(3.3)
	c	0.2	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.4(0.0)	0.4(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	0.9(0.0)	1.0(0.0)
r=9	r	9	9.2(1.6)	9.1(1.6)	9.1(0.7)	9.1(0.7)	24.0(9.6)	24.3(10.3)	23.6(5.9)	23.7(5.8)
	c	0.2	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.01)	0.2(0.0)	0.2(0.0)	0.4(0.0)	0.4(0.0)
	σ	1	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	0.99(0.1)	0.9(0.0)	1.0(0.0)

Table 3.5 Number of variables selected (Exponential covariance, p=36)

	PMLE		PREG		FAIR	
p=36	N-s	N-c	N-s	N-c	N-s	N-c
r=1	21(6.2)	9(1.7)	19(7.0)	10(0.8)	6(4.0)	5(2.4)
r=2	20(5.9)	9(1.7)	19(6.8)	10(0.9)	5(3.3)	4(1.9)
r=3	19(5.6)	9(1.8)	20(7.7)	10(0.8)	4(2.4)	3(1.7)
r=4	20(5.6)	9(1.7)	20(7.4)	10(0.4)	3(1.6)	3(1.5)
r=5	20(5.5)	10(1.5)	20(7.1)	10(0)	3(1.5)	3(1.5)
r=6	20(5.2)	10(1.2)	20(7.4)	10(0.1)	3(1.4)	3(1.4)
r=7	20(5.1)	10(1.2)	19(7.3)	10(0)	3(1.4)	3(1.4)
r=8	20(4.8)	10(1.0)	20(7.7)	10(0)	3(1.4)	3(1.4)
r=9	20(5.0)	10(1.1)	19(7.6)	10(0)	3(1.4)	3(1.3)

Key: N-s: the number of variables selected by the model; N-c: the number of correct variables selected by the model.

Table 3.6 Number of variables selected (Exponential covariance, p=400)

	PMLE		PREG		$PMLE_T$		$PREG_T$		FAIR	
p=400	N-s	N-c	N-s	N-c	N-s	N-c	N-s	N-c	N-s	N-c
r=1	85(55.3)	9(1.5)	42(55.2)	9(1.4)	82(51.3)	10(1)	46(64.7)	9(1.4)	21(15.1)	7(2.1)
r=2	92(39.3)	10(1.4)	56(77.8)	9(1.1)	90(38.9)	10(1.2)	44(62.3)	9(1.4)	20(15.6)	7(2.5)
r=3	82(31.5)	10(1.2)	50(69.1)	9(1.2)	80(37)	10(1.4)	51(76.9)	9(1.3)	18(14.7)	6(2.6)
r=4	80(35.7)	10(0.8)	53(68.6)	9(1.1)	72(29.7)	10(0.7)	55(79.9)	10(1.1)	14(11.7)	6(2.7)
r=5	74(32.2)	10(0.6)	51(73.3)	9(1.2)	65(32.2)	10(0.7)	65(97.5)	10(0.9)	11(10.6)	5(2.8)
r=6	67(27.8)	10(0.6)	64(91.7)	10(1.1)	58(24.6)	10(0.4)	64(96)	10(1.1)	9(9.2)	5(2.8)
r=7	59(20.5)	10(0.5)	56(91)	10(1.1)	56(35.1)	10(0.5)	64(91.4)	10(0.5)	8(8.5)	4(2.7)
r=8	57(22.3)	10(0.5)	61(99)	10(1.1)	49(28.9)	10(0.5)	62(85)	10(1.2)	8(8.2)	4(2.8)
r=9	53(21.9)	10(0.5)	64(98.5)	10(1.1)	47(35.7)	10(0.4)	65.(91.8)	10(0.7)	7(7.6)	4(2.7)

Key: N-s: the number of variables selected by the model; N-c: the number of correct variables selected by the model.

Table 3.7 Number of variables selected (Exponential covariance, p=1225)

	$PMLT_T$		$PREG_T$		FAIR	
p=1225	N-s	N-c	N-s	N-c	N-s	N-c
r=1	116(185.7)	8(2.1)	52(97.2)	9(1.8)	31(22.3)	7.0(1.7)
r=2	143(192.3)	9(2.2)	53(128.3)	9(1.7)	37(22.7)	7.0(2.1)
r=3	133(135.1)	9(1.8)	63(151.9)	9(1.7)	34(24.5)	6.9(2.4)
r=4	104(91.5)	9(1.3)	56(129.7)	9(1.7)	28(19.6)	6.5(2.8)
r=5	95(77.4)	10(1.0)	56(115.2)	9(1.7)	26(20.6)	6.4(2.8)
r=6	76(74.0)	10(0.8)	49(107.9)	9(2.0)	23(18.1)	6.2(3.0)
r=7	66(54.3)	10(0.7)	49(87.2)	9(1.8)	21(16.2)	6.1(3.1)
r=8	48(46.3)	10(0.6)	53(110.2)	9(1.6)	18(15.6)	5.6(3.3)
r=9	48(48.1)	10(0.7)	55(104.3)	9(1.7)	15(14.0)	5.3(3.4)

Key: N-s: the number of variables selected by the model; N-c: the number of correct variables selected by the model.

Table 3.8 Classification Accuracy Rate (Polynomial covariance, p=36)

	TURE	MLE	PREG	PMLE	FAIR	NB
ρ			p=36			
0	0.94(0.02)	0.92(0.02)	0.92(0.03)	0.92(0.02)	0.89(0.04)	0.92(0.04)
0.1	0.92(0.02)	0.89(0.03)	0.89(0.03)	0.89(0.03)	0.87(0.04)	0.89(0.04)
0.2	0.90(0.02)	0.86(0.03)	0.86(0.04)	0.86(0.04)	0.84(0.04)	0.87(0.04)
0.3	0.88(0.02)	0.84(0.03)	0.84(0.04)	0.84(0.05)	0.81(0.04)	0.84(0.04)
0.4	0.88(0.02)	0.83(0.03)	0.83(0.03)	0.83(0.05)	0.78(0.04)	0.81(0.04)
0.5	0.88(0.02)	0.83(0.03)	0.83(0.04)	0.84(0.04)	0.76(0.05)	0.79(0.05)
0.6	0.89(0.02)	0.85(0.03)	0.85(0.04)	0.85(0.04)	0.74(0.04)	0.77(0.04)
0.7	0.91(0.02)	0.88(0.02)	0.88(0.03)	0.88(0.04)	0.72(0.05)	0.75(0.05)
0.8	0.94(0.02)	0.92(0.02)	0.92(0.03)	0.92(0.03)	0.71(0.05)	0.74(0.05)
0.9	0.99(0.01)	0.98(0.01)	0.97(0.02)	0.97(0.02)	0.70(0.05)	0.73(0.05)

Table 3.9 Classification Accuracy Rate (Polynomial covariance, p=400)

	TURE	MLE	MLE_T	PREG	$PREG_T$	PMLE	$PMLE_T$	FAIR	NB
ρ				p=400					
0	0.94(0.02)	0.79(0.03)	0.79(0.03)	0.89(0.04)	0.89(0.04)	0.90(0.03)	0.90(0.03)	0.88(0.04)	0.79(0.03)
0.1	0.93(0.02)	0.76(0.03)	0.77(0.03)	0.87(0.03)	0.87(0.04)	0.86(0.04)	0.87(0.04)	0.87(0.04)	0.78(0.03)
0.2	0.92(0.02)	0.75(0.03)	0.76(0.03)	0.85(0.04)	0.85(0.04)	0.83(0.05)	0.84(0.04)	0.86(0.04)	0.77(0.03)
0.3	0.91(0.02)	0.74(0.03)	0.74(0.03)	0.84(0.05)	0.85(0.05)	0.82(0.05)	0.84(0.04)	0.83(0.04)	0.74(0.04)
0.4	0.92(0.02)	0.74(0.03)	0.75(0.03)	0.83(0.05)	0.84(0.05)	0.83(0.05)	0.84(0.04)	0.82(0.04)	0.71(0.04)
0.5	0.92(0.02)	0.75(0.03)	0.76(0.03)	0.84(0.05)	0.84(0.05)	0.85(0.05)	0.86(0.05)	0.80(0.04)	0.68(0.04)
0.6	0.93(0.02)	0.77(0.03)	0.78(0.03)	0.86(0.05)	0.86(0.04)	0.88(0.04)	0.89(0.04)	0.78(0.04)	0.65(0.04)
0.7	0.95(0.01)	0.81(0.03)	0.81(0.03)	0.88(0.05)	0.88(0.05)	0.92(0.03)	0.92(0.03)	0.75(0.04)	0.61(0.05)
0.8	0.97(0.01)	0.87(0.02)	0.87(0.02)	0.93(0.04)	0.93(0.04)	0.96(0.03)	0.96(0.03)	0.73(0.04)	0.59(0.05)
0.9	1.0(0)	0.97(0.01)	0.97(0.01)	0.98(0.02)	0.98(0.02)	0.99(0.02)	0.99(0.02)	0.70(0.06)	0.56(0.05)

Table 3.10 Classification Accuracy Rate (Polynomial covariance, p=1225)

	TURE	MLE_T	$PREG_T$	$PMLE_T$	FAIR	NB
ρ			p=1225			
0	0.94(0.02)	0.70(0.04)	0.88(0.04)	0.88(0.04)	0.88(0.05)	0.70(0.05)
0.1	0.93(0.02)	0.68(0.04)	0.86(0.05)	0.85(0.05)	0.86(0.04)	0.69(0.04)
0.2	0.92(0.02)	0.67(0.03)	0.83(0.05)	0.81(0.06)	0.85(0.05)	0.68(0.05)
0.3	0.91(0.02)	0.67(0.03)	0.83(0.04)	0.80(0.06)	0.83(0.04)	0.66(0.04)
0.4	0.91(0.02)	0.67(0.03)	0.82(0.05)	0.80(0.06)	0.81(0.04)	0.64(0.04)
0.5	0.92(0.02)	0.68(0.03)	0.83(0.06)	0.83(0.05)	0.79(0.04)	0.61(0.04)
0.6	0.93(0.02)	0.69(0.03)	0.84(0.05)	0.86(0.04)	0.76(0.05)	0.59(0.05)
0.7	0.95(0.02)	0.71(0.03)	0.86(0.06)	0.89(0.04)	0.74(0.06)	0.57(0.06)
0.8	0.97(0.01)	0.75(0.04)	0.89(0.07)	0.94(0.02)	0.71(0.07)	0.55(0.07)
0.9	1.00(0)	0.85(0.04)	0.95(0.04)	0.98(0.02)	0.69(0.07)	0.53(0.07)

Table 3.11 Parameter estimation (Polynomial covariance)

	p=36		p=400				p=1225	
TURE	MLE	PMLE	MLE	MLE_T	PMLE	$PMLE_T$	MLE_T	$PMLE_T$
$\rho = 0$	0.0(0.0)	0.0(0.0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
$\sigma = 1$	1.0(0.0)	1(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0)	1.0(0.0)
$\rho = 0.1$	0.1(0.0)	0.1(0.0)	0.1(0.0)	0.1(0.0)	0.1(0.1)	0.1(0.0)	0.1(0.0)	0.1(0.0)
$\sigma = 1$	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)
$\rho = 0.2$	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.2(0.0)	0.1(0.1)	0.2(0.0)	0.2(0.0)	0.2(0.0)
$\sigma = 1$	1.0(0.0)	1.0(0.0)	1.0(0.0)	1(0.0)	1.0(0.0)	1(0.0)	1.0(0.0)	1.0(0.0)
$\rho = 0.3$	0.3(0.0)	0.3(0.0)	0.3(0.0)	0.3(0.0)	0.3(0.0)	0.3(0.0)	0.2(0)	0.2(0)
$\sigma = 1$	1.0(0.0)	1(0.0)	1.0(0.01)	1(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1(0.0)
$\rho = 0.4$	0.4(0.0)	0.4(0.0)	0.4(0.0)	0.4(0.0)	0.4(0.0)	0.4(0.0)	0.3(0.0)	0.3(0.0)
$\sigma = 1$	1.0(0.0)	1(0.0)	1.0(0.0)	1(0.0)	1.0(0.0)	1(0.0)	1.0(0.0)	1.0(0.1)
$\rho = 0.5$	0.5(0.0)	0.5(0.0)	0.5(0.0)	0.5(0.0)	0.5(0.0)	0.5(0.0)	0.3(0)	0.3(0)
$\sigma = 1$	1.0(0.0)	1(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1(0.0)	1.0(0.0)	1.0(0.0)
$\rho = 0.6$	0.6(0.0)	0.6(0.0)	0.6(0.0)	0.6(0.0)	0.6(0.0)	0.6(0.0)	0.4(0.0)	0.4(0.0)
$\sigma = 1$	1.0(0.1)	1(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)
$\rho = 0.7$	0.7(0.0)	0.7(0.0)	0.7(0.0)	0.7(0.01)	0.7(0.0)	0.7(0.0)	0.5(0.0)	0.5(0.0)
$\sigma = 1$	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.0)	1.0(0.0)	0.9(0.0)	1.0(0.0)
$\rho = 0.8$	0.8(0.0)	0.8(0.0)	0.8(0.0)	0.8(0.0)	0.8(0.0)	0.8(0.0)	0.5(0.0)	0.5(0.0)
$\sigma = 1$	1.0(0.1)	1.0(0.1)	1.0(0.0)	1.0(0.0)	1.0(0.1)	1.0(0.1)	0.9(0.0)	0.9(0.0)
$\rho = 0.9$	0.9(0.0)	0.9(0.01)	0.9(0.0)	0.9(0.01)	0.9(0.0)	0.9(0.0)	0.7(0.0)	0.7(0.0)
σ=1	1.0(0.1)	1.1(0.1)	1.0(0.1)	1.0(0.1)	1.0(0.1)	1.1(0.1)	0.8(0.0)	0.9(0.0)

Table 3.12 Number of variables selected (Polynomial covariance, p=36)

p=36	PMLE		PREG		FAIR	
ρ	N-s	N-c	N-s	N-c	N-s	N-c
0	20(5.0)	10(0.4)	17(7.2)	10(0.7)	9(3.0)	7(2.0)
0.1	21(5.4)	10(0.7)	18(6.5)	10(0.6)	8(3.2)	7(2.3)
0.2	20(5.2)	10(1.4)	18(6.4)	10(0.4)	7(3.7)	6(2.5)
0.3	21(6.2)	10(1.8)	19(7.0)	10(0.8)	7(4.3)	6(2.4)
0.4	20(5.9)	10(1.8)	20(7.0)	10(0.8)	6(4.1)	5(2.4)
0.5	21(5.6)	10(1.7)	20(6.9)	10(0.7)	5(3.9)	4(2.2)
0.6	21(5.6)	10(1.6)	21(7.3)	10(0.7)	4(2.9)	4(2)
0.7	20(5.9)	10(1.9)	20(7.3)	10(0.2)	3(2.4)	3(1.8)
0.8	19(5.6)	9(2.2)	19(7.7)	10(0)	3(1.5)	3(1.5)
0.9	15(5.9)	7(2.8)	15(7.9)	10(0.3)	2(1.1)	2(1.0)

Key: N-s: the number of variables selected by the model; N-c: the number of correct variables selected by the model.

Table 3.13 Number of variables selected (Polynomial covariance, p=400)

p=400	PMLE		PREG		$PMLE_T$		$PREG_T$		FAIR	
ρ	N-s	N-c	N-s	N-c	N-s	N-c	N-s	N-c	N-s	N-c
0	46(47.1)	9(1)	38.5(49.8)	9(1.3)	47.4(47.4)	9.4(0.9)	38(49.6)	9(1.3)	15(12.1)	7(1.9)
0.1	51(51.9)	9(1.2)	41(50.4)	9(1.1)	52(57.6)	9(1.1)	36(46.7)	9.1(1.2)	17(13.1)	7(1.9)
0.2	77(57.3)	9(1.4)	35(54.5)	9(1.5)	73(60.6)	9(1.3)	39(58.5)	9(1.5)	18(12.8)	7(1.9)
0.3	84(50.5)	9(1.5)	40(51.9)	9(1.4)	82(48.9)	9(1.2)	39(56.4)	9(1.3)	19(13.9)	7(2.1)
0.4	97(47.3)	10(1.5)	59(75.3)	9(1.3)	88(43.3)	10(0.9)	52(70.3)	9(1.3)	22(15.7)	7(2.2)
0.5	93(42.7)	10(1.5)	55(78.7)	9(1.1)	92(42.6)	10(1.2)	46(57.1)	9(1.2)	22(15.6)	7(2.4)
0.6	85(32.4)	10(1.3)	52(75.2)	9(1.2)	83(31.2)	10(1)	47(63.2)	9(1.2)	21(15.8)	7(2.4)
0.7	78(26.4)	10(0.7)	54(69.4)	10(1.2)	72(24.9)	10(0.7)	59(78.6)	10(1.3)	16(13)	6(2.8)
0.8	60(27.9)	10(1.2)	68(95.6)	10(0.5)	56(27.5)	10(1.1)	69(96)	10(0.5)	10(9.6)	5(3)
0.9	29(30)	9(1.5)	55(97.1)	10(1.2)	29(24.8)	9(1.6)	56(98.1)	10(1.2)	5(5.8)	3(2.6)

Key: N-s: the number of variables selected by the model; N-c: the number of correct variables selected by the model.

Table 3.14 Number of variables selected (Polynomial covariance, p=1225)

p=1225	$PMLT_T$		$PREG_T$		FAIR	
ρ	N-s	N-c	N-s	N-c	N-s	N-c
0	43(67.9)	9(1.4)	35(48.0)	9(1.6)	17(13.2)	7(1.8)
0.1	53(73.5)	8(1.7)	35(49.1)	8(1.8)	24(17.7)	7(1.7)
0.2	112(136.7)	8(2.2)	45(70.1)	8(1.8)	27(19.0)	7(1.7)
0.3	121(147.5)	8(2.3)	52(87.2)	9(1.6)	31(19.4)	7(1.8)
0.4	166(160.5)	8(2.5)	58(123.7)	9(1.7)	34(19.8)	7(1.9)
0.5	152(99.5)	9(1.6)	49(74.9)	9(2.1)	36(21.6)	7(2.0)
0.6	140(91.2)	9(1.4)	58(96.0)	9(2.1)	38(23.5)	7(2.3)
0.7	113(82.5)	9(1.5)	79(184.8)	9(1.8)	31(22.2)	7(2.8)
0.8	87(77.6)	10(0.7)	82(171.0)	9(2.0)	23(19.7)	6(3.3)
0.9	25(40.2)	10(1.1)	76(173.7)	9(1.8)	12(11.8)	5(3.6)

Key: N-s: the number of variables selected by the model; N-c: the number of correct variables selected by the model.

Table 3.15 Classification Accuracy Rate (Mis-specified covariance, p=36)

	TURE	MLE	PREG	PMLE	FAIR	NB
	p=36					
r=1	0.887(0.02)	0.841(0.03)	0.84(0.04)	0.843(0.05)	0.826(0.05)	0.854(0.05)
r=2	0.879(0.02)	0.851(0.03)	0.85(0.04)	0.848(0.04)	0.755(0.05)	0.771(0.05)
r=3	0.904(0.02)	0.888(0.03)	0.883(0.03)	0.889(0.03)	0.727(0.06)	0.746(0.06)
r=4	0.925(0.02)	0.907(0.02)	0.906(0.03)	0.91(0.03)	0.713(0.05)	0.737(0.05)
r=5	0.939(0.02)	0.922(0.02)	0.922(0.02)	0.925(0.03)	0.709(0.05)	0.735(0.05)
r=6	0.951(0.02)	0.935(0.02)	0.935(0.02)	0.935(0.02)	0.701(0.06)	0.734(0.06)
r=7	0.958(0.01)	0.943(0.02)	0.943(0.02)	0.946(0.02)	0.7(0.05)	0.735(0.05)
r=8	0.963(0.01)	0.95(0.02)	0.949(0.02)	0.953(0.02)	0.701(0.05)	0.736(0.05)
r=9	0.969(0.01)	0.956(0.01)	0.954(0.02)	0.959(0.02)	0.705(0.05)	0.736(0.05)

Table 3.16 Classification Accuracy Rate (Mis-specified covariance, p=400)

	TURE	MLE	MLE_T	PREG	$PREG_T$	PMLE	$PMLE_T$	FAIR	NB
	p=400								
r=1	0.91(0.02)	0.73(0.03)	0.74(0.03)	0.84(0.05)	0.84(0.05)	0.83(0.05)	0.84(0.04)	0.85(0.05)	0.75(0.03)
r=2	0.93(0.02)	0.76(0.03)	0.78(0.03)	0.86(0.04)	0.87(0.04)	0.89(0.04)	0.90(0.04)	0.79(0.04)	0.67(0.04)
r=3	0.95(0.01)	0.82(0.03)	0.82(0.03)	0.89(0.04)	0.90(0.04)	0.93(0.03)	0.93(0.03)	0.76(0.04)	0.62(0.04)
r=4	0.97(0.01)	0.85(0.02)	0.86(0.02)	0.92(0.03)	0.93(0.03)	0.96(0.02)	0.96(0.02)	0.74(0.04)	0.60(0.04)
r=5	0.98(0.01)	0.88(0.02)	0.88(0.02)	0.94(0.03)	0.94(0.03)	0.97(0.02)	0.96(0.02)	0.73(0.04)	0.59(0.04)
r=6	0.98(0.01)	0.90(0.02)	0.90(0.02)	0.95(0.03)	0.95(0.03)	0.97(0.02)	0.97(0.01)	0.71(0.05)	0.58(0.04)
r=7	0.99(0.01)	0.92(0.02)	0.87(0.06)	0.96(0.03)	0.94(0.04)	0.98(0.01)	0.96(0.02)	0.71(0.07)	0.57(0.04)
r=8	0.99(0.01)	0.93(0.02)	0.69(0.11)	0.96(0.03)	0.86(0.07)	0.98(0.01)	0.88(0.05)	0.70(0.07)	0.56(0.04)
r=9	1.00(0.01)	0.94(0.02)	0.61(0.05)	0.97(0.02)	0.84(0.06)	0.98(0.01)	0.87(0.03)	0.70(0.06)	0.56(0.04)

Table 3.17 Classification Accuracy Rate (Mis-specified covariance, p=1225)

	TURE	MLE_T	$PREG_T$	$PMLE_T$	FAIR	NB
	p=1225					
r=1	0.91(0.02)	0.66(0.04)	0.83(0.04)	0.81(0.05)	0.83(0.05)	0.67(0.05)
r=2	0.93(0.02)	0.69(0.03)	0.84(0.05)	0.88(0.04)	0.78(0.05)	0.60(0.05)
r=3	0.95(0.02)	0.72(0.03)	0.87(0.05)	0.92(0.04)	0.75(0.04)	0.57(0.04)
r=4	0.97(0.01)	0.74(0.03)	0.89(0.06)	0.94(0.02)	0.73(0.05)	0.55(0.05)
r=5	0.98(0.01)	0.75(0.04)	0.90(0.06)	0.95(0.02)	0.73(0.05)	0.54(0.05)
r=6	0.98(0.01)	0.76(0.04)	0.91(0.06)	0.95(0.02)	0.72(0.05)	0.54(0.05)
r=7	0.99(0.01)	0.75(0.04)	0.91(0.06)	0.95(0.02)	0.712(0.05)	0.54(0.05)
r=8	0.99(0.01)	0.72(0.05)	0.90(0.07)	0.94(0.02)	0.71(0.06)	0.54(0.06)
r=9	0.99(0.01)	0.65(0.07)	0.86(0.09)	0.90(0.05)	0.71(0.06)	0.53(0.06)

3.5 Real Data Analysis

Data used in the real data example were obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (http:// www.loni.ucla.edu/ADNI), which is lauched in 2004, aiming to improve clinical trials for prevention and treatment of Alzheimer's disease (AD). In the interest of promoting consistency in data analysis, the ADNI Core has created standardized analysis sets of the structure MRI scans comprising only image data that have passed quality control (QC) assessments conducted at the Aging and Dementia Imaging Research laboratory at the Mayo Clinic (see [27]). In this study, we used T1-weighted MRI images from the collection of standardized datasets. The description of the standardized MRI imaging from ADNI can be found in `http://adni.loni.usc.edu/methods/mri-analysis/adni-standardized-data/` and [54].

According to [27], the images were obtained using magnetization prepared rapid gradient echo (MPRAGE) or equivalent protocols with varying resolutions (typically 1.0 $\times$ 1.0 mm in plane spatial resolution and 1.2 mm thick sagittal slices with 256*256*166 voxels). The images were then pre-processed according to a number of steps detailed in [27] and `http://adni.loni.usc.edu/methods/mri-analysis/mri-pre-processing/`, which corrected gradient non-linearity, intensity inhomogeneity and phantom-based distortion. In addition, the pre-processed imaging were processed by FreeSurfer for cortical reconstruction and volumetric segmentation by Center for Imaging of Neurodegnerative Diseases, UCSF. The skull-stripped volume (brain mask) obtained by FreeSurfer cross-sectional processing were used in this study.

Only images from ADNI-1 subjects obtained using 1.5 T scanners at screening visits were used in this study, and we used the first time point if there are multiple images of the

Table 3.18 Subjects characteristics

	AD	NL	p-value
n	187	227	
Age (Mean±sd)	75.28 ± 7.55	75.80 ± 4.98	0.4168
Gender (F/M)	88/99	110/117	0.813
MMSE (Mean±sd)	23.28 ± 2.04	29.11 ± 1.00	$< 1e-15$

Key: AD, subjects with Alzheimer's disease ; NL, healthy subjects; Age, baseline age; MMSE, baseline Mini-Mental State Examination.

same subject acquired at different times. 187 subjects diagnosed as Alzheimer's disease at screening visits and 227 healthy subjects at screening visits are contained in this study. The total number of subjects is 414. Details of the subjects can be found in Table 3.18.

After downloading the pre-processed imaging data from ADNI, an R package ANTsR were applied for imaging registration. Then we use "3dresample" command by AFNI software ([16]) to adjust the resolution and reduce the total number of voxels of the imaging to 18*22*18. Take x axis and y axis for horizontal plane, x axis and z axis for coronal plane and y axis and z axis for sagittal plane. Only the 1100 voxels located in the center of the brain were used as features for classification (i.e. the voxels with x coordinate from 5 to 14; y coordinate from 6 to 16; z coordinate from 5 to 14).

After removing the voxels with zero signal for most of the subjects (more than 409 subjects), we have 1077 voxels left in use. The distance between each pair of voxels can be calculated by their coordinates. For example, there are two voxels s_1, s_2 with coordinate $s_1 = (x_1, y_1, z_1)$ and $s_2 = (x_2, y_2, z_2))$. Then the distance between s_1 and s_2 is defined by: $d(s_1, s_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$. From the plot of sample covariance, spatial correlations can be observed among voxels.

We random sample 100 from the 187 AD subjects and 100 from the 227 health subjects as the training set. Then there are 87 AD subjects and 127 health subjects left. The testing

Table 3.19 Subjects characteristics of training and testing set

		training set	testing set	p-value
AD	n	100	87	
	Age (Mean±sd)	75.64 ± 7.39	74.85 ± 7.75	0.478
	Gender (F/M)	47/53	41/46	0.999
	MMSE (Mean±sd)	23.22 ± 2.08	23.36 ± 2.01	0.649
NL	n	100	87	
	Age (Mean±sd)	75.99 ± 5.39	75.34 ± 4.56	0.3723
	Gender (F/M)	42/58	50/37	0.05
	MMSE (Mean±sd)	29.06 ± 1.04	29.09 ± 1.01	0.8307

Key: AD, subjects with Alzheimer's disease ; NL, healthy subjects; Age, baseline age; MMSE, baseline Mini-Mental State Examination.

Table 3.20 Classification performance for voxel level MRI data. Training and testing samples are of sizes 200 and 174, respectively

	MLE	PREG	PMLE	FAIR	NB
Accuracy	0.707	0.747	0.776	0.621	0.689
No. of training err	38	47	46	77	55
No. of testing err	51	44	39	66	54
No. of selected voxels	1077	3	5	7	1077

set includes the 87 AD subjects and a random sample of 87 from the 127 health subjects. Detail of the subjects in the training and testing set can be found in Table 3.19.

We assume the exponential correlation among voxels. Then we apply the PLDA method proposed in this research for classification. First, the parameter are estimated by PMLE: $r = 6.59, c = 0.924, \sigma^2 = 230.38$ and 5 voxels are selected for classification meanwhile from training data. Then we plug in the estimates in the classification function and do classification on the testing data.

The classification accuracy rate of PLDA is listed in Table 3.20. We also listed the classification accuracy from other methods. It shows the classification accuracy rate of our method is about 77.6%, which is superior to other comparable methods (MLE-LDA: 70.7%, PREG-LDA: 74.7%, FAIR: 62.1% and NB: 68.9%).

3.6 Proofs of the main results

3.6.1 Proofs for classification using MLE

Lemma 3.6.1. *Let $\boldsymbol{\epsilon}$ be p-dimensional vectors and $\boldsymbol{\epsilon} \sim N(0, \Sigma)$, where Σ is a $p \times p$ positive definite covariance matrix. For m-dimension vector $\boldsymbol{u}$ with $\|\boldsymbol{u}\|_2 = \sqrt{\boldsymbol{u}^T\boldsymbol{u}} = C$ and $p \times m$ matrix $\boldsymbol{X}_i$, we have:*

$$\left|\boldsymbol{\epsilon}^T\boldsymbol{X}u\right| = O_p(\sqrt{tr(\boldsymbol{X}^T\Sigma\boldsymbol{X})}\,\|\boldsymbol{u}\|_2) \tag{3.6.1}$$

Proof. Since $E(\boldsymbol{\epsilon}^T\boldsymbol{X}) = 0$,

$$\begin{aligned} E(\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{u})^2 \leq \left[E(\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\epsilon})\right]^{1/2}\|\boldsymbol{u}\|_2 &= \left[E(tr(\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{\epsilon}))\right]^{1/2}\|\boldsymbol{u}\|_2 \\ =tr(\boldsymbol{X}^T\Sigma\boldsymbol{X})\,\|\boldsymbol{u}\|_2 & \end{aligned} \tag{3.6.2}$$

By Chebyshev's inequality, for any M

$$P(\frac{\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{u}}{\sqrt{\|\boldsymbol{u}\|_2^2\,tr(\boldsymbol{X}^T\Sigma\boldsymbol{X})}} > M) \leq \frac{E(\boldsymbol{\epsilon}^T\boldsymbol{X}u)^2}{M^2\,\|\boldsymbol{u}\|_2^2\,tr(\boldsymbol{X}^T\Sigma\boldsymbol{X})} = \frac{1}{M^2} \tag{3.6.3}$$

Thus for any $\epsilon > 0$, exits M large enough such that

$$P(\frac{\left|\boldsymbol{\epsilon}^T\boldsymbol{X}\boldsymbol{u}\right|}{\sqrt{\|\boldsymbol{u}\|_2^2\,tr(\boldsymbol{X}^T\Sigma\boldsymbol{X})}} > M) < \epsilon \tag{3.6.4}$$

This lead to $\left|\boldsymbol{\epsilon}^T\boldsymbol{X}u\right| = O_p(\sqrt{tr(\boldsymbol{X}^T\Sigma\boldsymbol{X})}\,\|\boldsymbol{u}\|_2)$. □

Lemma 3.6.2. *Let $\boldsymbol{\epsilon}_i (i = 1, 2, ..., n)$ be p-dimensional vectors and $\boldsymbol{\epsilon}_i \sim N(0, c(n)\Sigma)$, where*

$c(n)$ is a function of n and Σ is a $p \times p$ positive definite covariance matrix with $\lambda(\Sigma) < \infty$. For a $p \times p$ matrix A, we have

$$\sum_{i=1}^{n} \left[\boldsymbol{\epsilon}_i^T A \boldsymbol{\epsilon}_i - c(n) tr(A\Sigma) \right] = O_p(c(n)\sqrt{n} \left\| A \right\|_F)$$

Proof. Since $E(\boldsymbol{\epsilon}_i^T A \boldsymbol{\epsilon}_i) = tr(c(n)A\Sigma)$, we have

$$E(\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^T A \boldsymbol{\epsilon}_i - c(n)tr(A\Sigma))^2 = \sum_{i=1}^{n} E(\boldsymbol{\epsilon}_i^T A \boldsymbol{\epsilon}_i - c(n)tr(A\Sigma))^2 = \sum_{i=1}^{n} E(\boldsymbol{\epsilon}_i^T A \boldsymbol{\epsilon}_i)^2 - nc^2(n)tr^2(A\Sigma) \tag{3.6.5}$$

Let $B = c(n)\Sigma^{\frac{1}{2}} A \Sigma^{\frac{1}{2}}$, then exit orthogonal matrix Q such that $B = Q^T \Lambda Q$ where $\Lambda = diag(\lambda_i)$ and λ_i are eigenvalues of B. Let $\tilde{\boldsymbol{\epsilon}}_i = \sqrt{c(n) Q \Sigma^{-\frac{1}{2}}} \boldsymbol{\epsilon}_i$, then $\tilde{\boldsymbol{\epsilon}}_i \sim N(0, I_{p\times p})$ where $I_{p\times p}$ is identity matrix. Then

$$\begin{aligned} E(\boldsymbol{\epsilon}_i^T A \boldsymbol{\epsilon}_i)^2 &= E(\tilde{\boldsymbol{\epsilon}}_i^T \Lambda \tilde{\boldsymbol{\epsilon}}_i)^2 = E(\sum_{i=1}^{p} \lambda_j \tilde{\boldsymbol{\epsilon}}_{ij}^2)^2 \\ &= E(\sum_{i=1}^{p} \lambda_j^2 \tilde{\boldsymbol{\epsilon}}_{ij}^4 + \sum_{j,k=1}^{p} \lambda_j \lambda_k \tilde{\boldsymbol{\epsilon}}_{ij}^2 \tilde{\boldsymbol{\epsilon}}_{ij}^2) = 2\sum_{i=1}^{p} \lambda_j^2 + (\sum_{i=1}^{p} \lambda_j)^2 \\ &= 2tr(B^T B) + tr^2(B) \\ &= c(n)^2 [2tr(\Sigma A \Sigma A) + tr^2(A\Sigma)] \end{aligned} \tag{3.6.6}$$

Hence

$$E(\sum_{i=1}^{n} \boldsymbol{\epsilon}_i^T A \boldsymbol{\epsilon}_i - c(n)tr(A\Sigma))^2 = 2nc(n)^2 tr(\Sigma A^T \Sigma A) \leq 2nc(n)^2 \lambda_{\max}^2(\Sigma) tr(A^T A) \tag{3.6.7}$$

By Chebyshev's inequality, for any M we have

$$P(\frac{\sum_{i=1}^{n}\boldsymbol{\epsilon}_i^T A\boldsymbol{\epsilon}_i - c(n)tr(A\Sigma)}{\sqrt{nc(n)^2\left\|A\right\|_F^2}} > M) \leq \frac{2nc(n)^2 tr(\Sigma A^T\Sigma A)}{M^2 nc(n)^2\left\|A\right\|_F^2} \leq \frac{2\lambda_{\max}^2(\Sigma)}{M^2} \tag{3.6.8}$$

Then for any $\epsilon > 0$ exits M large enough such that

$$p(\frac{\sum_{i=1}^{n}\boldsymbol{\epsilon}_i^T A\boldsymbol{\epsilon}_i - c(n)tr(A\Sigma)}{\sqrt{nc(n)^2\left\|A\right\|_F^2}}) > M) < \epsilon \tag{3.6.9}$$

which means $\sum_{i=1}^{n}\boldsymbol{\epsilon}_i^T A\boldsymbol{\epsilon}_i - c(n)tr(A\Sigma) = O_p(c(n)\sqrt{n}\left\|A\right\|_F)$ □

Proof of Theorem 1. Take derivative with respect to $\boldsymbol{\mu}_k$ $(k = 1, 2)$ with the function $L(\boldsymbol{\theta}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ defined in 3.1.6. Considering $\Sigma^{-1}(\boldsymbol{\theta})$ is nonsingular, we have $\boldsymbol{\mu}_k = \hat{\boldsymbol{\mu}}_{kMLE} = \bar{\boldsymbol{Y}}_{k\cdot}$.

For $\hat{\boldsymbol{\theta}}_{MLE}$, we first consider the case of $p/n \to 0$. It is sufficient to prove that for any given $\epsilon > 0$, there is a large constant C such that for large p and n, the smallest rate of convergence $\eta_{n,p}$ is $\sqrt{\frac{1}{np}}$ such that we have

$$P(\sup_{\left\|\boldsymbol{u}\right\|_2 = C} L(\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) < L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2)) > 1 - \epsilon \tag{3.6.10}$$

where $\boldsymbol{u} \in \mathbb{R}^q$. This implies that there exists a local maximum for the function L in the neighborhood of $\boldsymbol{\theta}_0$ with the radius at most proportional to $\eta_{n,p}$.

$$\begin{aligned} &L(\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) - L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) \\ &= (\frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0))^T \boldsymbol{u}\eta_{n,p} + \tfrac{1}{2}\boldsymbol{u}^T(\frac{\partial^2 L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\boldsymbol{\theta}^*))\boldsymbol{u}\eta_{n,p}^2 \\ &= -\frac{n}{2}\boldsymbol{u}^T T(\boldsymbol{\theta}_0)\boldsymbol{u}\eta_{n,p}^2 + (\frac{\partial L}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0))^T \boldsymbol{u}\eta_{n,p} + \tfrac{1}{2}\boldsymbol{u}^T(\frac{\partial^2 L}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}(\boldsymbol{\theta}^*) + mnT(\boldsymbol{\theta}_0))\boldsymbol{u}\eta_{n,p}^2 \end{aligned} \tag{3.6.11}$$

$$= (I) + (II) + (III)$$

where $T(\boldsymbol{\theta}_0)$ is a $q \times q$ matrix with its $(i,j)th$ element $t_{ij}(\boldsymbol{\theta}_0) = tr(\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_j)$.

From A4 and A5, $t_{ij} = a_{ij}(t_{ii})^{\frac{1}{2}}(t_{jj})^{\frac{1}{2}} \geq a_{ij}\lambda_{\min}^{-2}(\Sigma)\left\|\Sigma_i\right\|_F \left\|\Sigma_j\right\|_F$. There exists a constant M such that

$$(I) = -\frac{n}{2}\boldsymbol{u}^T T\boldsymbol{u}\eta_{n,p}^2 = -\frac{n}{2}\sum_{i,j=1}^{q} t_{ij}u_iu_j\eta_{n,p}^2 \leq -\frac{Mnp}{2}\eta_{n,p}^2\left\|\boldsymbol{u}\right\|_2^2 \tag{3.6.12}$$

and

$$\begin{aligned}(II) &= -\tfrac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}\sum_{j=1}^{q}\left[(\boldsymbol{Y}_{ki}-\hat{\boldsymbol{\mu}}_k)^T\Sigma^j(\boldsymbol{Y}_{ki}-\hat{\boldsymbol{\mu}}_k) - (\frac{n_k-1}{n_k})tr(\Sigma\Sigma^j)\right]u_{\theta_j}\eta_{n,p} \\ &+ \tfrac{1}{2}\sum_{j=1}^{q} tr(\Sigma\Sigma^j)u_{\theta_j}\eta_{n,p} \\ &= (1) + (2)\end{aligned} \tag{3.6.13}$$

Because $\boldsymbol{Y}_{ki} - \hat{\boldsymbol{\mu}}_k \sim N(0, \frac{n_k-1}{n_k}\Sigma)$, by lemma 3.6.2,

$$|(1)| = O_p(\sum_{k=1}^{2}\frac{n_k-1}{\sqrt{n_k}}\left\|\Sigma^j\right\|_F\left\|\boldsymbol{u}\right\|_2\eta_{n,p}) = O_p(\sqrt{n}\left\|\Sigma^j\right\|_F\left\|\boldsymbol{u}\right\|_2\eta_{n,p}) = O_p(\sqrt{np}\left\|\boldsymbol{u}\right\|_2\eta_{n,p}) \tag{3.6.14}$$

The last equality is because from A6(i), $\left\|\Sigma^i\right\|_F^2 \leq \left\|\Sigma^i\right\|_2^2 = p\lambda_{\max}^2(\Sigma^i) = O(p)$.

Also by A6(i), $tr(\Sigma\Sigma^j) = tr(\Sigma^{1/2}\Sigma^j\Sigma^{1/2}) \leq \lambda_{\max}(\Sigma^j)tr(\Sigma)$. Then noticing that $tr(\Sigma) =$

$O(p)$, and $p/n \to 0$

$$|(2)| = \left| \tfrac{1}{2} \sum_{j=1}^{q} tr(\Sigma\Sigma^j) u_j \eta_{n,p} \right| = O_p(tr(\Sigma)\eta_{n,p} \left\|\boldsymbol{u}\right\|_2) = O_p(p\eta_{n,p} \left\|\boldsymbol{u}\right\|_2)$$

Thus

$$(II) = O_p((\sqrt{np} + p)\eta_{n,p} \left\|\boldsymbol{u}\right\|_2)$$

- If $p/n \to 0$, $(II) = O_p(\sqrt{np})\eta_{n,p}$. By choosing sufficient large $C = \left\|\boldsymbol{u}\right\|_2$, the minimal rate of $\eta_{n,p}$ to have (II) be dominated by (I) is $\eta_{n,p} = O_p(\sqrt{\frac{1}{np}})$;
- If $p/n \to C$ with $0 < C \leq \infty$, $(II) = O_p(p\eta_{n,p})$. Then the minimal rate of $\eta_{n,p}$ to have (II) dominated by (I) is $\eta_{n,p} = O_p(\sqrt{\frac{1}{n}})$

Since

$$\frac{\partial L}{\partial \theta_j \partial \theta_l}(\boldsymbol{\theta}) = \frac{n}{2}\left[tr(\Sigma^j(\boldsymbol{\theta})\Sigma(\boldsymbol{\theta})) + tr(\Sigma^j(\boldsymbol{\theta})\Sigma_l(\boldsymbol{\theta}))\right] - \frac{1}{2}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(\boldsymbol{Y}_{ki} - \hat{\boldsymbol{\mu}}_k)^T\Sigma^{jl}(\boldsymbol{\theta})(\boldsymbol{Y}_{ki} - \hat{\boldsymbol{\mu}}_k)$$

(III) can be written as

$$\begin{aligned}
&- \tfrac{1}{2}\sum_{j,l=1}^{q}\left[\sum_{k=1}^{2}\sum_{i=1}^{n_k}\left((\boldsymbol{Y}_{ki} - \hat{\boldsymbol{\mu}}_k)^T\Sigma^{jl}(\boldsymbol{\theta}^*)(\boldsymbol{Y}_{ki} - \hat{\boldsymbol{\mu}}_k) - (\frac{n_k - 1}{n_k})tr(\Sigma(\boldsymbol{\theta}_0)\Sigma^{jl}(\boldsymbol{\theta}^*))\right)\right] u_{\theta_j} u_{\theta_l} \eta_{n,p}^2 \\
&+ \frac{n}{2n_1 n_2}\sum_{j,l=1}^{q} tr(\Sigma^{jl}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}_0)) u_{\theta_j} u_{\theta_l} \eta_{n,p}^2 \\
&+ \frac{n}{2}\sum_{j,l=1}^{q}\left[tr(\Sigma^{jk}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}^*)) - tr(\Sigma^{jk}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}_0))\right] u_{\theta_j} u_{\theta_l} \eta_{n,p}^2 \\
&+ \frac{n}{2}\sum_{j,l=1}^{q}\left[tr(\Sigma^{j}(\boldsymbol{\theta}^*)\Sigma_l(\boldsymbol{\theta}^*)) - tr(\Sigma^{j}(\boldsymbol{\theta}_0)\Sigma_l(\boldsymbol{\theta}_0))\right] u_{\theta_j} u_{\theta_l} \eta_{n,p}^2
\end{aligned}$$

$=(3)+(4)+(5)+(6)$

By lemma 3.6.2 and A6(ii),

$$|(3)| = O_p(\sqrt{n}\left\|\Sigma^{jl}(\theta^*)\right\|_F \eta_{n,p}^2) = O_p(\sqrt{np}\eta_{n,p}^2)$$

For (4), by A6(ii)

$$tr(\Sigma^{jl}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}_0)) = O_p(p)$$

thus $|(4)| = O_p(\frac{p}{n}\eta_{n,p}^2)$ It is easy to see (3), (4) are dominated by (I). For (5),

$$|(5)| = \left|\frac{1}{4}n\sum_{j,k=1}^{q} tr(\Sigma^{kj}(\boldsymbol{\theta}^*)(\Sigma(\boldsymbol{\theta}_0)-\Sigma(\boldsymbol{\theta}^*)))u_{\theta_k}u_{\theta_j}\eta_{n,p}^2\right|. \tag{3.6.15}$$

Let $d_{il}(\boldsymbol{\theta}^*)$ be the i,lth entry of matrix $\Sigma^{kj}(\boldsymbol{\theta}^*)$, $\gamma_{il}(\boldsymbol{\theta})$ be the i,lth entry of $\Sigma(\boldsymbol{\theta})$, then by A2 and A6($ii$)

$$\begin{aligned} tr(\Sigma^{kj}(\boldsymbol{\theta}^*)(\Sigma(\boldsymbol{\theta}_0)-\Sigma(\boldsymbol{\theta}^*))) &= \sum_{i,l=1}^{p} d_{il}(\boldsymbol{\theta}^*)(\gamma_{li}(\boldsymbol{\theta}_0)-\gamma_{li}(\boldsymbol{\theta}^*)) \\ &\leq \sum_{i,l=1}^{p} |d_{il}(\boldsymbol{\theta}^*)|\| \frac{\partial\gamma_{il}(\theta^*)}{\partial\boldsymbol{\theta}} \|_2 \| \boldsymbol{\theta}_0-\boldsymbol{\theta}^* \|_2 \\ &\leq \sum_{i,l=1}^{p} d_{il}(\boldsymbol{\theta}_2^*)M\eta_{n,p} \\ &\leq M\eta_{n,p}p\left\|\Sigma^{kj}(\boldsymbol{\theta}^*)\right\|_F \\ &= O_p(\sqrt{p^3}\eta_{n,p}) \end{aligned} \tag{3.6.16}$$

Hence

$$|(5)| = O_P(n\sqrt{p^3}\eta^3) = O_p((p^{3/2}n\eta_{n,p})\eta_{n,p}^2)) \tag{3.6.17}$$

- If $p/n \to 0$ and $\eta_{n,p} = O_p(\frac{1}{\sqrt{np}})$, (5) is dominated by (I).
- If $p/n \to C$ with $0 < C \leq \infty$, $eta_{n,p} = O_p(\frac{1}{n})$ and $\sqrt{p}/n \to 0$, (5) is dominated by (I).

For (6), let $t_{ij}(\theta^*) = tr(\Sigma^{-1}(\boldsymbol{\theta}^*)\Sigma_i(\boldsymbol{\theta}^*)\Sigma^{-1}(\boldsymbol{\theta}^*)\Sigma_j(\boldsymbol{\theta}^*))$, by A6$(iii)$,

$$\begin{aligned} |(6)| &\leq n \sum_{k,j=1}^{q} \left\| \frac{\partial t_{ij}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \right\|_2 \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 \, u_{\theta_k} u_{\theta_j} \eta_{n,p}^2 \\ &= O_P(np\eta_{n,p}^3) \end{aligned} \tag{3.6.18}$$

While $\eta_{n,p} = O_p(\frac{1}{\sqrt{np}})$ or $\eta_{n,p} = O_p(\frac{1}{n})$, (6) is also dominated by (I). Hence (III) is dominated by (I). This completes the proof. □

Proof of Theorem 2. We start with $W_1(\hat{\delta}_{MLE}; \Theta) = 1 - \Phi(\Psi_1)$, where $W_1(\hat{\delta}_{MLE}; \Theta)$ is the conditional misclassification rate defined in 1.3.2 and Ψ_1 is defined in 1.3.3. The idea is to prove $\liminf_{n,p\to 0} \Psi_1 \to \frac{\sqrt{C_0}}{2}$.

From Therom 1, we have $\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \|_2 = O_p(\frac{1}{\sqrt{np}})$. Recall that $\Sigma = \Sigma(\boldsymbol{\theta}) = \left[\gamma(h_{ij}; \boldsymbol{\theta})\right]_{i,j=1}^{p}$ and $\hat{\Sigma} = \Sigma(\hat{\boldsymbol{\theta}}) = \left[\gamma(h_{ij}; \hat{\boldsymbol{\theta}})\right]_{i,j=1}^{p}$.

By A2, we have:

$$\max_{i,j} |\gamma(h_{ij}; \boldsymbol{\theta}) - \gamma(h_{ij}; \hat{\boldsymbol{\theta}})| \leq M \| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \|_2 \tag{3.6.19}$$

Thus there exist $\epsilon > 0$ and matrix $E = \left[e_{ij}\right]_{i,j=1}^{p}$ such that

$$\hat{\Sigma} = \Sigma + \epsilon E \tag{3.6.20}$$

where $\epsilon = O_p(\frac{1}{\sqrt{np}})$ and E is a $p \times p$ matrix with absolute values of all entries less than 1, i.e. $\left|e_{ij}\right| \leq 1$ for any $i, j = 1, 2, ..., p$. As a result, for large p and n, the inverse of Σ can be written as:

$$\hat{\Sigma}^{-1} = \Sigma^{-1} - \epsilon\Sigma^{-1}E\Sigma^{-1} + O(\epsilon^2)E_2 \tag{3.6.21}$$

where $\mathbb{E}_2$ is a $p \times p$ matrix with all entries less than 1, see [39].

Now we consider the denominator of 1.3.3. We first claim the denominator can be written as:

$$\hat{\mathbf{\Delta}}^T(\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1})\hat{\mathbf{\Delta}} = \hat{\mathbf{\Delta}}^T\Sigma^{-1}\hat{\mathbf{\Delta}}(1 + o_p(1)). \tag{3.6.22}$$

Because by 3.6.21, we have

$$\begin{aligned}\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1} &= (\Sigma^{-1} - \epsilon\Sigma^{-1}E\Sigma^{-1} + O(\epsilon^2)E_2)\Sigma(\Sigma^{-1} - \epsilon\Sigma^{-1}E\Sigma^{-1} + O(\epsilon^2)E_2) \\ &= \Sigma^{-1} - 2\epsilon A + \epsilon^2 A\Sigma A + O(\epsilon^2)E_2 + O(\epsilon^3)EE_2\Sigma^{-1} + O(\epsilon^4)E_2E_2\end{aligned} \tag{3.6.23}$$

where $A = \Sigma^{-1}E\Sigma^{-1}$.

Also, noticing that $\epsilon = O(\frac{1}{\sqrt{np}})$, $k_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq k_2$ and $\lambda_{\max}(E) \leq tr(E) \leq$

p, we have:

$$\frac{\hat{\boldsymbol{\Delta}}^T(\epsilon A)\hat{\boldsymbol{\Delta}}}{\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}} = \frac{y^T\epsilon E y}{y^T\Sigma y} \leq \frac{\epsilon\lambda_{\max}(E)}{\lambda_{\min}(\Sigma)} \leq \epsilon p/\lambda_{\min}(\Sigma) = O(\sqrt{\frac{p}{n}}) \tag{3.6.24}$$

where $y^T = \hat{\boldsymbol{\Delta}}^T\Sigma^{-1}$. Similarly, we have:

$$\frac{\hat{\boldsymbol{\Delta}}^T(\epsilon^2 A\Sigma A)\hat{\boldsymbol{\Delta}}}{\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}} \leq \epsilon^2\frac{\lambda^2_{\max}(E)}{\lambda^2_{\min}(\Sigma)} \leq \epsilon^2 p^2/\lambda^2_{\min}(\Sigma) = O(\frac{p}{n}) \tag{3.6.25}$$

$$\frac{\hat{\boldsymbol{\Delta}}^T(O(\epsilon^2)E_2)\hat{\boldsymbol{\Delta}}}{\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}} \leq O(\epsilon^2)\lambda_{\max}(E_2)\lambda_{\min}(\Sigma) \leq O(\epsilon^2)p\lambda_{\max}(\Sigma) = O(\frac{1}{n}) \tag{3.6.26}$$

$$\frac{\hat{\boldsymbol{\Delta}}^T(O(\epsilon^3)EE_2\Sigma^{-1})\hat{\boldsymbol{\Delta}}}{\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}} \leq O(\epsilon^3)\frac{\lambda_{\max}(E_2)\lambda_{\max}(E)\lambda_{\max}(\Sigma^{-1})}{\lambda_{\min}(\Sigma)} \leq O(\epsilon^3)p^2\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} = O(\sqrt{\frac{p}{n}}\frac{1}{n}) \tag{3.6.27}$$

$$\frac{\hat{\boldsymbol{\Delta}}^T(O(\epsilon^4)E_2E_2)\hat{\boldsymbol{\Delta}}}{\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}} \leq O(\epsilon^4)p^2\lambda_{\max}(\Sigma) = O_p(\frac{1}{n^2}) \tag{3.6.28}$$

Since $\frac{p}{n} \to 0$ as $n \to \infty$ and $p \to \infty$, (3.6.22) is derived by combining (3.6.24)- (3.6.28).

Now we investigate $\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}$ and claim that:

$$\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}} = \Delta^T\Sigma^{-1}\Delta(1 + o_p(1)) + \frac{np}{n_1n_2}(1 + o_p(1)) \tag{3.6.29}$$

Recall $\hat{\boldsymbol{\mu}}_1 = \bar{\boldsymbol{Y}}_{1\cdot} = \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{Y}_{1i}$, which is normally distributed as $\mathcal{N}(\boldsymbol{\mu}_1, \frac{1}{n_1}\Sigma)$. Also, $\hat{\boldsymbol{\mu}}_2 = \bar{\boldsymbol{Y}}_{2\cdot} = \frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{Y}_{2i}$, which is normally distributed as $N(\boldsymbol{\mu}_2, \frac{1}{n_2}\Sigma)$. Let $\hat{\boldsymbol{\mu}}_1 = \boldsymbol{\mu}_1 + \hat{\boldsymbol{\epsilon}}_1$

and $\hat{\boldsymbol{\mu}}_2 = \boldsymbol{\mu}_2 + \hat{\boldsymbol{\epsilon}}_2$ where $\hat{\boldsymbol{\epsilon}}_1 \sim \mathcal{N}(0, \frac{1}{n_1}\Sigma)$ and $\hat{\boldsymbol{\epsilon}}_2 \sim N(0, \frac{1}{n_2}\Sigma)$. Then we have:

$$\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta} + 2\Delta^T\Sigma^{-1}(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2) + (\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)^T\Sigma^{-1}(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)$$

Noticing $\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2 \sim N(0, \frac{n}{n_1 n_2}\Sigma)$, by Chebyshev's inequality, for any $\epsilon_0 > 0$

$$\begin{aligned} P(\frac{\boldsymbol{\Delta}^T\Sigma^{-1}(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)}{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}} > \epsilon_0) &\leq \frac{E\big(\boldsymbol{\Delta}^T\Sigma^{-1}(\hat{\epsilon}_1 - \hat{\epsilon}_2)\big)^2}{(\epsilon_0\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta})^2} \\ &= \frac{n}{n_1 n_2 \boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}} \\ &\leq \frac{n}{n_1 n_2 C_p} = \frac{1}{\pi(1-\pi)nC_p} \to 0 \end{aligned} \tag{3.6.30}$$

It goes to 0 because $nC_p \to \infty$. Then

$$\boldsymbol{\Delta}^T\Sigma^{-1}(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2) = o_p(\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta})$$

Then we consider the third term in 3.6.29. Let $\tilde{\boldsymbol{\epsilon}} = \sqrt{\frac{n_1 n_2}{n}}\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)$. Then $\tilde{\boldsymbol{\epsilon}} \sim \mathcal{N}(0, I_{p\times p})$. Now for any $\varepsilon_0 > 0$

$$\begin{aligned} P(|\frac{(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)^T\Sigma^{-1}(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2) - np/n_1 n_2}{np/n_1 n_2}| > \epsilon_0) &= P(|\frac{\tilde{\boldsymbol{\epsilon}}^T\tilde{\boldsymbol{\epsilon}} - p}{p}| > \epsilon_0) \\ &\leq \frac{E\big(\tilde{\boldsymbol{\epsilon}}^T\tilde{\boldsymbol{\epsilon}}\big)^2}{\epsilon_0^2 p^2} \\ &= \frac{2}{p}\frac{1}{\epsilon^2} \to 0 \end{aligned}$$

as $p \to \infty$. Then

$$(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2)^T\Sigma^{-1}(\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2) = \frac{np}{n_1 n_2}(1 + o_p(1)) \tag{3.6.31}$$

Then 3.6.29 followed. Now 3.6.22 and 3.6.29 yield:

$$\hat{\boldsymbol{\Delta}}^T(\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1})\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}^T\Sigma\boldsymbol{\Delta}(1+o_p(1)) + \frac{np}{n_1n_2}(1+o_p(1)) \tag{3.6.32}$$

Now we consider the nominator of 1.3.3.

$$\begin{aligned}(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}})^T\hat{\Sigma}\hat{\boldsymbol{\Delta}} =& \frac{1}{2}\Big(\boldsymbol{\Delta}^T\hat{\Sigma}^{-1}\boldsymbol{\Delta} + \hat{\boldsymbol{\epsilon}}_2^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\epsilon}}_2 - \hat{\boldsymbol{\epsilon}}_1^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\epsilon}}_1 - 2\boldsymbol{\Delta}^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\epsilon}}_2\Big) \\ =& \frac{1}{2}((1)+(2)-(3)-(4))\end{aligned} \tag{3.6.33}$$

$$(1) = \boldsymbol{\Delta}^T(\Sigma^{-1} - \epsilon E + O(\epsilon^2)E_2)\boldsymbol{\Delta}$$

By the assumption that $k_1 \le \lambda_{\min}(\Sigma) \le \lambda_{\max} \le k_2$, $\lambda_{\max}(E) \le p$ and $\lambda_{\max}(E_2) \le p$, we have

$$\frac{\boldsymbol{\Delta}^T(\epsilon E)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}} \to 0$$

and

$$\frac{\boldsymbol{\Delta}^T(O(\epsilon^2)E_2)\boldsymbol{\Delta}}{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}} \to 0$$

thus

$$(1) = \boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1+o_p(1)) \tag{3.6.34}$$

By the same argument, we have $(2) = \hat{\boldsymbol{\epsilon}}_2^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_2(1+o_p(1))$ and $(3) = \hat{\boldsymbol{\epsilon}}_1^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_1(1+o_p(1))$.

Since $\hat{\boldsymbol{\epsilon}}_1 \sim N(0, \frac{1}{n_1}\Sigma)$ and $\hat{\boldsymbol{\epsilon}}_1 \sim N(0, \frac{1}{n_1}\Sigma)$,similar to the proof of (3.6.31), we have:

$$(2) = \frac{p}{n_2}(1 + o_p(1)) \text{ and } (3) = \frac{p}{n_1}(1 + o_p(1)) \tag{3.6.35}$$

Now we consider term (4) in 3.6.33.

$$\begin{aligned}(4) =& \boldsymbol{\Delta}^T(\Sigma^{-1} - \epsilon\Sigma^{-1}E\Sigma^{-1} + O(\epsilon^2)E_2)\hat{\boldsymbol{\epsilon}}_2 \\ =& \boldsymbol{\Delta}^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_2 + \epsilon\boldsymbol{\Delta}^T\Sigma^{-1}E\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_2 + O(\epsilon^2)\boldsymbol{\Delta}^T E_2\hat{\boldsymbol{\epsilon}}_2\end{aligned}$$

Similar to the proof of 3.6.30, all the three terms in (4) are small order of $\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}$. Thus we have

$$(4) = o_p(\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}) \tag{3.6.36}$$

Now the nominator can be written as:

$$(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}})^T\hat{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} = \tfrac{1}{2}\left(\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1 + o_p(1)) + \frac{p}{n_1 n_2}(n_1 - n_2)(1 + o_p(1))\right) \tag{3.6.37}$$

3.6.32 and 3.6.37 yield

$$W_1(\hat{\delta}_{MLE}; \Theta) = 1 - \Phi\Big(\frac{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1 + o_p(1)) + \frac{p}{n_1 n_2}(n_1 - n_2)(1 + o_p(1))}{2\sqrt{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1 + o_p(1)) + \frac{np}{n_1 n_2}(1 + o_p(1))}}\Big)$$

By the same argument, we have:

$$W_2(\hat{\delta}_{MLE}; \Theta) = \Phi\Big(\frac{-\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1 + o_p(1)) + \frac{p}{n_1 n_2}(n_1 - n_2)(1 + o_p(1))}{2\sqrt{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1 + o_p(1)) + \frac{np}{n_1 n_2}(1 + o_p(1))}}\Big)$$

Since $\frac{p}{n} \to 0$ and $C_p = \boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta} \to C_0$ with $0 \leq C_0 \leq \infty$, we have

$$\begin{aligned} W(\hat{\delta}_{MLE}; \Theta) =& \frac{1}{2}\Bigg(1 - \Phi\left(\frac{C_p(1+o_p(1)) + \frac{p}{n_1 n_2}(n_1 - n_2)(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{np}{n_1 n_2}(1+o_p(1))}}\right) \\ &+ \Phi\left(\frac{-C_p(1+o_p(1)) + \frac{p}{n_1 n_2}(n_1 - n_2)(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{np}{n_1 n_2}(1+o_p(1))}}\right)\Bigg) \\ \to& 1 - \Phi(\frac{\sqrt{C_0}}{2}) \end{aligned}$$

as $p \to \infty$ and $n \to \infty$. If $C_p \to C_0 < \infty$, $1 - \Phi(\frac{\sqrt{C_0}}{2}) > 0$. Thus $\hat{\delta}_{MLE}$ is asymptotically optimal. Now we check the asymptotically optimal when $C_p \to \infty$. From the inequality

$$\frac{x}{1+x^2} e^{-\frac{x^2}{2}} \leq \Phi(-x) \leq \frac{1}{x} e^{-\frac{x^2}{2}}, \ x > 0 \tag{3.6.38}$$

we have

$$\frac{xy}{1+x^2} e^{-\frac{x^2-y^2}{2}} \leq \frac{W(\hat{\delta}_{MLE})}{\Phi(-\frac{\sqrt{C_p}}{2})} \leq \frac{1+y^2}{xy} e^{-\frac{x^2-y^2}{2}}$$

where $x = \frac{C_p(1+o_p(1)) \pm \frac{p}{n_1 n_2}(n_1 - n_2)(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{np}{n_1 n_2}(1+o_p(1))}}$ and $y = \frac{\sqrt{C_p}}{2}$. It is easy to check that $\frac{xy}{1+x^2} \to 1$ and $\frac{1+y^2}{xy} \to 1$ as $C_p \to \infty$. Also $x^2 - y^2 \to 0$ if $C_p(p/n) \to 0$. This completes the proof. $\square$

Proof of Theorem 3. The misclassification rate of $\delta_{\hat{\boldsymbol{\mu}}}$ is:

$$W(\delta_{\hat{\boldsymbol{\mu}}}; \Theta) = \frac{1}{2}(W_1(\delta_{\hat{\boldsymbol{\mu}}}; \Theta) + W_2(\delta_{\hat{\boldsymbol{\mu}}}; \Theta))$$

where

$$W_1(\delta_{\hat{\boldsymbol{\mu}}};\Theta) = 1 - \Phi(\Psi_1) \text{ and } W_2(\delta_{\hat{\boldsymbol{\mu}}};\Theta) = \Phi(\Psi_2)$$

where Ψ_1 and Ψ_2 is defined by 1.3.3 and 1.3.5 with $\hat{\Sigma}$ replaced by Σ. We start with

$$\Psi_1 = \frac{(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}})^T\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}{\sqrt{(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T\Sigma^{-1}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)}}$$

The denominator is (3.6.29) and it can be represented as:

$$\hat{\boldsymbol{\Delta}}\Sigma^{-1}\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}^T\Sigma\boldsymbol{\Delta}(1 + o_p(1)) + \frac{np}{n_1n_2}(1 + o_p(1))$$

The nominator is:

$$(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}})^T\Sigma^{-1}\hat{\boldsymbol{\Delta}} = \tfrac{1}{2}(\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta} + \hat{\boldsymbol{\epsilon}}_1^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_2 - 2\boldsymbol{\Delta}^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_2)$$

By the similar procedure in the proof of (3.6.37), it can be represented by

$$(\boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}})^T\Sigma^{-1}\hat{\boldsymbol{\Delta}} = \tfrac{1}{2}\left(\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1 + o_p(1)) + \frac{p}{n_1n_2}(n_1 - n_2)(1 + o_p(1))\right)$$

Thus we have:

$$W_1(\hat{\delta}_{\hat{\boldsymbol{\mu}}};\Theta) = 1 - \Phi\Big(\frac{C_p(1 + o_p(1)) + \frac{p}{n_1n_2}(n_1 - n_2)(1 + o_p(1))}{2\sqrt{C_p(1 + o_p(1)) + \frac{np}{n_1n_2}(1 + o_p(1))}}\Big) \tag{3.6.39}$$

Similarly, we have

$$W_2(\hat{\delta}_{\hat{\boldsymbol{\mu}}};\Theta) = \Phi\Big(\frac{-C_p(1+o_p(1)) + \frac{p}{n_1 n_2}(n_1-n_2)(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{np}{n_1 n_2}(1+o_p(1))}}\Big)$$

(i) If $\frac{C_p}{p/n} \to \infty$. Then

$$\begin{aligned}\frac{\pm C_p(1+o_p(1)) + \frac{p}{n_1 n_2}(n_1-n_2)(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{np}{n_1 n_2}(1+o_p(1))}} &= \frac{\pm C_p(1 \pm \frac{p}{n_1 n_2 C_p}(n_1-n_2)(1+o_p(1)))}{2\sqrt{C_p(1+\frac{np}{n_1 n_2 C_p}(1+o_p(1))}} \\ &= \frac{\pm\sqrt{C_p}(1 \pm \frac{p}{n_1 n_2 C_p}(n_1-n_2)(1+o_p(1)))}{2\sqrt{(1+\frac{np}{n_1 n_2 C_p}(1+o_p(1))}} \\ &\to \frac{\pm\sqrt{C_0}}{2}\end{aligned}$$

which yields $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \to 0$ since $\frac{p}{n} \to C$ with $0 < C < \infty$ and $C_p \to C_0 = \infty$. Now we show that $\frac{W(\hat{\delta}_{\hat{\boldsymbol{\mu}}})}{W_{OPT}} \to \infty$ in probability.

Noticing the fact that

$$\frac{x}{1+x^2}e^{-\frac{x^2}{2}} \le \Phi(-x) \le \frac{1}{x}e^{-\frac{x^2}{2}},\ x > 0$$

we have

$$\frac{W_{OPT}}{\Phi\left(-\frac{x}{2}\right)} = \frac{\Phi(-\frac{\sqrt{C_p}}{2})}{\Phi\left(-\frac{x}{2}\right)} \le \frac{4+x^2}{x\sqrt{C_p}}e^{-\frac{1}{8}(C_p - x^2)}$$

where $x = \frac{C_p(1+o_p(1)\pm\frac{p(n_1-n_2)}{n_1n_2}(1+o_p(1)))}{\sqrt{C_p(1+o_p(1))+\frac{np}{n_1n_2}(1+o_p(1))}}$

$$\frac{4+x^2}{x\sqrt{C_p}} = \frac{4}{x\sqrt{C_p}} + \frac{x}{\sqrt{C_p}} \to \text{a constant}$$

because

$$\frac{1}{x\sqrt{C_p}} \to 0$$

and

$$\frac{x}{\sqrt{C_p}} = \begin{cases} \to 1 & \text{if } c = \infty, \\ \to a_0 & \text{if } c < \infty \end{cases}$$

where $a_0 = \frac{\sqrt{c}+1/\sqrt{c}}{\sqrt{c+1/(\pi(1-\pi))}}$. Also

$$C_p - x^2 = \frac{C_p^2 o_p(1) + C_p\frac{(n\pm(n_1-n_2))p}{n_1n_2}(1+o_p(1)) + \frac{p^2(n_1-n_2)^2}{n_1^2n_2^2}}{C_p(1+o_p(1)) + \frac{np}{n_1n_2}(1+o_p(1))} \to \infty$$

Thus we have

$$\frac{\Phi(-\frac{\sqrt{C_p}}{2})}{\Phi\left(-\frac{x}{2}\right)} \to 0 \tag{3.6.40}$$

As a result

$$\frac{W(\hat{\delta}_{\hat{\boldsymbol{\mu}}})}{W_{OPT}} \to \infty$$

(ii)While $\frac{C_p}{p/n} \to c$ with $0 < c < \infty$

$$\begin{aligned}\frac{\pm C_p(1+o_p(1)) + \frac{p}{n_1 n_2}(n_1-n_2)(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{np}{n_1 n_2}(1+o_p(1))}} &= \frac{\pm C_p(1 \pm \frac{p}{n_1 n_2 C_p}(n_1-n_2)(1+o_p(1)))}{2\sqrt{C_p(1+\frac{np}{n_1 n_2 C_p}(1+o_p(1))}} \\ &= \frac{\pm\sqrt{C_p}(1 \pm \frac{p}{n_1 n_2 C_p}(n_1-n_2)(1+o_p(1)))}{2\sqrt{(1+\frac{np}{n_1 n_2 C_p}(1+o_p(1))}} \\ &\to \frac{\pm\sqrt{C_0}(1 \pm \frac{1}{c}\frac{2\pi-1}{\pi(1-\pi)})}{2\sqrt{1+\frac{1}{c}\frac{1}{\pi(1-\pi)}}}\end{aligned}$$

Since $C_p \to C_0$,

$$W_1(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \to 1 - \Phi(\frac{\sqrt{C_0}(1+\frac{1}{c}\frac{2\pi-1}{\pi(1-\pi)})}{2\sqrt{1+\frac{1}{c}\frac{1}{\pi(1-\pi)}}})$$

and

$$W_2(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \to 1 - \Phi(\frac{\sqrt{C_0}(1-\frac{1}{c}\frac{2\pi-1}{\pi(1-\pi)})}{2\sqrt{1+\frac{1}{c}\frac{1}{\pi(1-\pi)}}})$$

If $\frac{p}{n} \to C$ with $0 < C < \infty$, then $0 < C_0 < \infty$. Since $\Phi(x)$ is convex function in the sense that $\frac{1}{2}(\Phi(x+\epsilon)+\Phi(x-\epsilon)) \le \Phi(x)$ for any $x > 0$ and $x > \epsilon > 0$,

$$lim_P W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) = lim_P \tfrac{1}{2}(W_1(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) + W_2(\hat{\delta}_{\hat{\boldsymbol{\mu}}})) \ge 1 - \Phi(\frac{\sqrt{C_0}}{2\sqrt{1+\frac{1}{c}\frac{1}{\pi(1-\pi)}}}) > 1 - \Phi(\frac{\sqrt{C_0}}{2})$$

where lim_P means converge in probability with $p \to \infty$ and $n \to \infty$.

If $\frac{p}{n} \to \infty$, then $C_0 = \infty$. Hence $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) = \frac{1}{2}(W_1(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) + W_2(\hat{\delta}_{\hat{\boldsymbol{\mu}}})) \to 0$. By similar argument in (i), we have $\frac{W(\hat{\delta}_{\hat{\boldsymbol{\mu}}})}{W_{OPT}} \to \infty$.

(iii)While $\frac{Cp}{p/n} \to 0$,

$$\frac{\pm C_p(1+o_p(1)) + \frac{p}{n_1 n_2}(n_1-n_2)(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{np}{n_1 n_2}(1+o_p(1))}} = \frac{\sqrt{\frac{p}{n}}(\pm C_p/(\frac{p}{n}) + \frac{n(n_1-n_2)}{n_1 n_2}(1+o_p(1)))}{2\sqrt{(C_p/(\frac{p}{n}) + \frac{n^2}{n_1 n_2}(1+o_p(1))}}$$

$$\begin{cases} \to \infty & \text{if } n_1 > n_2, \\ \to -\infty & \text{if } n_1 < n_2. \end{cases}$$

Which yields $W(\hat{\delta}_{\hat{\boldsymbol{\mu}}}) \to \frac{1}{2}$. □

Proof of Corollary 1. Noticing that when $n_1 = n_2 = n/2$, $\hat{\boldsymbol{\epsilon}}_1 \sim N(0, \frac{1}{n_1}\Sigma)$ and $\hat{\boldsymbol{\epsilon}}_2 \sim N(0, \frac{1}{n_1}\Sigma)$. Let $\tilde{\epsilon}_i = \sqrt{n_i}\Sigma^{\frac{1}{2}}\hat{\epsilon}_i$ for $i = 1, 2$. Then $\tilde{\epsilon}_i \sim N(0, I_p)$ and

$$E(\hat{\boldsymbol{\epsilon}}_1^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_2)^2 = E(\tilde{\epsilon}_1^T\tilde{\epsilon}_1 - \tilde{\epsilon}_2^T\tilde{\epsilon}_2)^2/N_1^2 = \sum_{j=1}^{p}(\tilde{\epsilon}_{1j}^2 - \tilde{\epsilon}_{2j}^2)^2/n_1^2 = 6p/n_1^2$$

Hence we have:

$$\hat{\boldsymbol{\epsilon}}_1^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_1 - \hat{\boldsymbol{\epsilon}}_2^T\Sigma^{-1}\hat{\boldsymbol{\epsilon}}_2 = O_p(\frac{\sqrt{p}}{n})$$

Similar to the proof of (3.6.39) we have:

$$W_1(\hat{\delta}_{\hat{\boldsymbol{\mu}}}; \Theta) \leq 1 - \Phi\Big(\frac{C_p(1+o_p(1)) + \frac{\sqrt{p}}{n}(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{4p}{n}(1+o_p(1))}}\Big)$$

and

$$W_2(\hat{\delta_{\boldsymbol{\mu}}}; \Theta) \leq \Phi\Big(\frac{-C_p(1+o_p(1)) + \frac{\sqrt{p}}{n}(1+o_p(1))}{2\sqrt{C_p(1+o_p(1)) + \frac{4p}{n}(1+o_p(1))}}\Big)$$

$$\frac{\pm C_p(1+o_p(1))+\frac{\sqrt{p}}{n}(1+o_p(1))}{2\sqrt{C_p(1+o_p(1))+\frac{4p}{n}(1+o_p(1))}}=\frac{\pm C_p/\sqrt{\frac{p}{n}}(1+o_p(1))+\frac{1}{\sqrt{n}}(1+o_p(1))}{2\sqrt{C_p/\frac{p}{n}(1+o_p(1))+4+o_p(1))}}$$

$$\begin{cases}\to\pm\infty & \text{if } \frac{C_p}{\sqrt{p/n}}\to\infty\\ \to\pm\frac{c}{4} & \text{if } \frac{C_p}{\sqrt{p/n}}\to c \text{ and } p/n\to\infty\\ \to\pm\frac{c}{2\sqrt{4+c/\sqrt{C}}} & \text{if } \frac{C_p}{\sqrt{p/n}}\to c \text{ and } p/n\to C<\infty\\ \to 0 & \text{if } \frac{C_p}{\sqrt{p/n}}\to 0\end{cases}$$

The proof of $\frac{W(\hat{\delta}_{\hat{\boldsymbol{\mu}}})}{W_{OPT}}\to\infty$ is the same as that in the proof of Theorem 3(1). This completes the proof.

□

3.6.2 Proofs for consistency of PMLE

Proof of Theorem 4. In the algorithm, we estimate $\boldsymbol{\beta}^{(0)}$ first. Then $\boldsymbol{\theta}^{(0)}$ is estimated by fixing $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(0)}$. Then update $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(1)}$ by fixing $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(0)}$. $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(1)}$ is updated in last step by fixing $\boldsymbol{\beta}=\boldsymbol{\beta}^{(1)}$. So the idea is to prove the theorem in the following sequence: (a) The consistency and sparsity of $\boldsymbol{\beta}^{(0)}$; (b) The consistency of $\boldsymbol{\theta}^{(0)}$; (c) The consistency and sparsity of $\boldsymbol{\beta}^{(1)}$; (d) The consistency of $\boldsymbol{\theta}^{(1)}$.

(a) We first prove $\left\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_0\right\|_2=O_p(\sqrt{\frac{s}{n}})$ and $\hat{\boldsymbol{\beta}}_2^{(0)}=0$ with probability tending to 1, where $\hat{\boldsymbol{\beta}}_2^{(0)}$ is the $p-s$ dimension sub-vector of $\hat{\boldsymbol{\beta}}^{(0)}=(\hat{\boldsymbol{\beta}}_1^{(0)T}\hat{\boldsymbol{\beta}}_2^{(0)T})^T$. The proof of (a) is the same as the proof of (c), except the loss function if defined as $R(\boldsymbol{\beta})$ which is negative of the penalized MLE function (3.3.1) with covariance matrix $\dot{\Sigma}$ replaced by $diag_{n-1}(I_p)$. Then the parameters are estimated by minimize the loss function. We

omit the proof here and illustrate the details in (c).

(b) Second we prove $\left\|\hat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}_0\right\|_2 = O_p(\sqrt{\frac{1}{np}})$. Write

$$F(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(0)}; \mathbf{Z}) = -\frac{np}{2} log(2\pi) - \tfrac{1}{2} log\left|\dot{\Sigma}(\boldsymbol{\theta})\right| - \tfrac{1}{2}(\mathbf{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})^T \dot{\Sigma}^{-1}(\boldsymbol{\theta})(\mathbf{Z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(0)})$$

It is sufficient to prove for any given $\epsilon > 0$ the smallest convergence rate of $\eta_{n,p}$ is $\sqrt{\frac{1}{np}}$ such that we have

$$P(\sup_{\|\boldsymbol{u}\|_= C} F(\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}, \hat{\boldsymbol{\beta}}^{(0)}; \mathbf{Z}) < F(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}^{(0)}; \mathbf{Z})) > 1 - \epsilon$$

This implies there exists a local maximum for the function $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}^{(0)}; \mathbf{Z})$ of $\boldsymbol{\theta}$ in the neighborhood of $\boldsymbol{\theta}_0$ with the radius at most proportional to $\eta_{n,p}$.

By Taylor's expansion, $\dot{\Sigma}(\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}) - \dot{\Sigma}(\boldsymbol{\theta}_0) = \sum_{j=1}^{q} \frac{\partial \dot{\Sigma}(\boldsymbol{\theta}^*)}{\partial \theta_j} u_{\theta_j} \eta_{n,p}$, where $\boldsymbol{\theta}^*$ is between $\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}$ and $\boldsymbol{\theta}_0$. Denote $\dot{\Sigma}^j(\boldsymbol{\theta}^*) = \frac{\partial \dot{\Sigma}(\boldsymbol{\theta}^*)}{\partial \theta_j}$, then

$$\begin{aligned}
&F(\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}, \hat{\boldsymbol{\beta}}^{(0)}; \mathbf{Z}) - F(\boldsymbol{\theta}_0, \hat{\boldsymbol{\beta}}^{(0)}; \mathbf{Z}) \\
=& \left[F(\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}, \boldsymbol{\beta}_0) - F(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\right] - \sum_{j=1}^{q} (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}_0)^T \dot{\Sigma}^j(\boldsymbol{\theta}^*) \boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0) u_{\theta_j} \eta_{n,p} \\
&- \tfrac{1}{2} \sum_{j=1}^{q} (\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0)^T X^T \dot{\Sigma}^j(\boldsymbol{\theta}^*) X(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0) u_{\theta_j} \eta_{n,p} \\
=& (I) + (II) + (III)
\end{aligned}$$

where

$$
\begin{aligned}
(I) =& F(\boldsymbol{\theta}_0 + \boldsymbol{u}\eta_{n,p}, \boldsymbol{\beta}_0) - F(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) \\
=& -\frac{n-1}{2}\boldsymbol{u}^T T \boldsymbol{u}\eta_{n,p}^2 + \left(\frac{\partial F}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_0)\right)^T \boldsymbol{u}\eta_{n,p} + \frac{1}{2}\boldsymbol{u}^T\left(\frac{\partial^2 F}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}^*) + (n-1)T\right)\boldsymbol{u}\eta_{n,p}^2 \\
=& (1) + (2) + (3) \\
(II) =& \sum_{j=1}^{q} (\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0)^T \boldsymbol{X}^T \dot{\Sigma}^j(\boldsymbol{\theta}^*)(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}_0) u_j \eta_{n,p} \\
(III) =& -\tfrac{1}{2}\sum_{j=1}^{q} (\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0)^T \boldsymbol{X}^T \dot{\Sigma}^j(\boldsymbol{\theta}^*)\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0) u_j \eta_{n,p}
\end{aligned}
$$

We consider (I) first. T in (1) is $q \times q$ matrix with its (i,j)th element as $t_{ij}(\boldsymbol{\theta}_0)$, where $t_{ij}(\boldsymbol{\theta}) = tr(\Sigma^{-1}(\boldsymbol{\theta})\Sigma_i(\boldsymbol{\theta})\Sigma^{-1}(\boldsymbol{\theta})\Sigma_j(\boldsymbol{\theta}))$. By the similar argument in proving the bound of (I) in theorem 1, there exist a constant K, such that $(1) = -\frac{n}{2}\sum_{i,j=1}^{q} t_{ij}(\boldsymbol{\theta}_0) u_i u_j \eta_{n,p}^2 \leq -Knp\eta_{n,p}^2 \left\|\boldsymbol{u}\right\|_2^2$ with probability tending to 1. In regarding to (2),

$$
\frac{\partial F}{\partial \theta_j}(\boldsymbol{\theta}_0) = \frac{n-1}{2} tr(\Sigma^j(\boldsymbol{\theta}_0)\Sigma(\boldsymbol{\theta}_0)) - \tfrac{1}{2}(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})^T \dot{\Sigma}^j(\boldsymbol{\theta}_0)(\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta})
$$

Notice that $\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta} \sim N(0, \dot{\Sigma}(\boldsymbol{\theta}))$ and $tr(\dot{\Sigma}^j\dot{\Sigma}) = (n-1)tr(\Sigma^j\Sigma)$. By lemma 3.6.2,

$$
\begin{aligned}
(2) =& O_p\left(\sqrt{tr(\dot{\Sigma}\dot{\Sigma}^j\dot{\Sigma}\dot{\Sigma}^j)}\eta_{n,p}\left\|\boldsymbol{u}\right\|_2\right) = O_p\left(\sqrt{(n-1)tr(\Sigma\Sigma^j\Sigma\Sigma^j)}\eta_{n,p}\left\|\boldsymbol{u}\right\|_2\right) \\
=& O_p\left(\sqrt{(n-1)\left\|\Sigma^j\right\|_F^2}\eta_{n,p}\left\|\boldsymbol{u}\right\|_2\right)\left\|\boldsymbol{u}\right\|_2
\end{aligned}
$$

By A3, A6(i), $(2) = O_p(\sqrt{np}\eta_{n,p})$.

Then we consider (3). For any $j,k=1,2,...,q$

$$\frac{\partial^2(F)}{\partial\theta_j\partial\theta_k}(\boldsymbol{\theta}^*)=\frac{n-1}{2}(tr(\Sigma^{jk}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}^*)-t_{jk}(\boldsymbol{\theta}^*))-\tfrac{1}{2}(\mathbf{Z}-\boldsymbol{X\beta})^T\dot{\Sigma}^{jk}(\boldsymbol{\theta}^*)(\mathbf{Z}-\boldsymbol{X\beta})$$

Thus (3) could be written as:

$$\begin{aligned}(3)=&\sum_{j,k=1}^{q}\frac{n-1}{2}(tr(\Sigma^{jk}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}_0))-\tfrac{1}{2}(\mathbf{Z}-\boldsymbol{X\beta})^T\Sigma^{jk}(\boldsymbol{\theta}^*)(\mathbf{Z}-\boldsymbol{X\beta}))u_{\theta_j}u_{\theta_k}\eta_{n,p}^2\\&+\sum_{j,k=1}^{q}\frac{n-1}{2}(tr(\Sigma^{jk}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}^*))-tr(\Sigma^{jk}(\boldsymbol{\theta}^*)\Sigma(\boldsymbol{\theta}_0)))u_{\theta_j}u_{\theta_k}\eta_{n,p}^2\\&+\sum_{j,k=1}^{q}\frac{n-1}{2}(t_{jk}(\boldsymbol{\theta}_0)-t_{jk}(\boldsymbol{\theta}^*))u_{\theta_j}u_{\theta_k}\eta_{n,p}^2\\=&(i)+(ii)+(iii)\end{aligned}$$

By lemma 3.6.2 and A6(ii), $(i)=O_p(\sqrt{(n-1)tr(\Sigma\Sigma^{jk}\Sigma\Sigma^{jk})}\eta_{n,p}^2)=O_p(\sqrt{np}\eta_{n,p}^2)$. Similar to the deriving the order of (5) and (6) in the proof of Theorem1, $(ii)=O_p(n\sqrt{p^3}\eta_{n,p}^3)$ and $(iii)=O_p(np\eta_{n,p}^3)$. By choosing large $C=\|\boldsymbol{u}\|_2$, the minimal rate that (2) and (3) are dominated by (1) is $\eta_{n,p}=O_p(\frac{1}{\sqrt{np}})$.

Now we consider (II). Denote $B=\sum_{i=1}^{n-1}\boldsymbol{X}_i$. Then by lemma 3.6.1, for any $j=1,2,...,q$,

$$(\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_0)^T\boldsymbol{X}^T\dot{\Sigma}^j(\boldsymbol{\theta}^*)(\mathbf{Z}-\boldsymbol{X\beta}_0)=O_p(\sqrt{tr(\boldsymbol{X}^T\dot{\Sigma}^j(\boldsymbol{\theta}^*)\dot{\Sigma}(\boldsymbol{\theta}^*)\dot{\Sigma}^j(\boldsymbol{\theta}^*)\boldsymbol{X})}\left\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_0\right\|_2)$$

$$=O_p(\sqrt{tr(\sum_{i=1}^{n-1}\boldsymbol{X}_j^T\Sigma^{j*}\Sigma^*\Sigma^{j*}\boldsymbol{X}_i)+tr(B^T\Sigma^{j*}\Sigma^*\Sigma^{j*}B)}\left\|\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_0\right\|_2)$$

where $\Sigma^*=\Sigma(\boldsymbol{\theta}^*)$ and $\Sigma^{j*}=\Sigma^j(\boldsymbol{\theta}^*)$. Notice that $\sum_{i=1}^{n-1}\boldsymbol{X}_i^T\boldsymbol{X}_i=(\frac{n_1n_2}{n}-\frac{n_1^2}{n^2})I_{p\times p}$

and $(\sum_{i=1}^{n-1} \boldsymbol{X}_i)^T(\sum_{i=1}^{n-1} \boldsymbol{X}_i) = \frac{n_1^2}{n^2} I_{p\times p}$. Then by A6($i$)

$$tr(\sum_{i=1}^{n-1} \boldsymbol{X}_j^T \Sigma^{j*}\Sigma^*\Sigma^{j*}\boldsymbol{X}_i) = (\frac{n_1 n_2}{n} - \frac{n_1^2}{n_2}) tr(\Sigma^{j*}\Sigma^*\Sigma^{j*}) \leq \lambda_{\max}(\Sigma)\frac{n_1 n_2}{n} \left\|\Sigma^{j*}\right\|_F^2 = O_p(np).$$

Similarly $tr(B^T\Sigma^{j*}\Sigma^*\Sigma^{j*}B) = O_p(np)$.

Since $\left\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}_0\right\|_2 = O_p(\sqrt{\frac{s}{n}})$, $(II) = O_p(\sqrt{ps}\eta_{n,p})$

For (III), by A6(i), for any $j = 1, 2, ..., q$ we also have:

$$\begin{aligned}
&(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0)^T \boldsymbol{X}^T \dot{\Sigma}^j(\boldsymbol{\theta}^*)\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0) \\
=&(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0)^T \boldsymbol{X}^T diag_{n-1}\Sigma^j(\boldsymbol{\theta}^*)(\tilde{I}_{n-1,p} + \tilde{J}_{n-1,p})\boldsymbol{X}(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0) \\
\leq&\lambda_{\max}(\Sigma^{j*})((\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0)^T \left(\sum_{i=1}^{n-1} \boldsymbol{X}_i^T\boldsymbol{X}_i + (\sum_{i=1}^{n-1} \boldsymbol{X}_i)^T(\sum_{i=1}^{n-1} \boldsymbol{X}_i)\right)(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0)) \\
=&O_p(n\left\|\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0\right\|_2^2) = O_p(s)
\end{aligned}$$

Thus $(III) = O_p(s\eta_{n,p})$. Both (II) and (III) are dominated by (I) while $\eta_{n,p} = O_p(\sqrt{\frac{1}{np}})$. This concludes the proof of $\left\|\hat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}_0\right\|_2 = O_p(\sqrt{\frac{1}{np}})$

(c) Write $\hat{\boldsymbol{\beta}}^{(1)} = (\hat{\boldsymbol{\beta}}_1^{(1)}, \hat{\boldsymbol{\beta}}_2^{(2)})^T$. Then we prove $\left\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0\right\|_2 = O_p(\sqrt{\frac{s}{n}})$ and $\hat{\boldsymbol{\beta}}_2^{(1)} = 0$ with probability tending to 1, where $\hat{\boldsymbol{\beta}}_{(1)}^1$ formed by elements in $supp(\boldsymbol{\beta}_0)$ and $\hat{\boldsymbol{\beta}}_2^{(1)}$ is a $p-s$ subvector of $\hat{\boldsymbol{\beta}}^{(1)}$. Let $\eta_{n,p} = O_p(\left\|\hat{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}_0\right\|_2) = O_p(\sqrt{\frac{1}{np}})$. We use two steps to prove the consistency and sparsity.

step 1 We first prove consistency on s-dimensional space. Define the loglikelihood

function for $\boldsymbol{\beta}^1$ as

$$\bar{Q}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\beta}^1) = -\frac{np}{2}log(2\pi) - \tfrac{1}{2}log\left|\dot{\Sigma}(\hat{\boldsymbol{\theta}}^{(0)})\right| - \tfrac{1}{2}(\mathbf{Z} - \boldsymbol{X}^1\boldsymbol{\beta}^1)^T\dot{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}^{(0)})(\mathbf{Z} - \boldsymbol{X}^1\boldsymbol{\beta}^1) - n\sum_{j=1}^{n} P_\lambda(\left|\beta_j\right|)$$

where $\boldsymbol{\beta}^1$ is subvector of $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}^{1T}, \boldsymbol{\beta}^{2T})^T$ formed by elements in supp $\boldsymbol{\beta}_0$. We first $\left\|\hat{\boldsymbol{\beta}}_1^{(1)} - \boldsymbol{\beta}^1\right\|_2 = O_p(\sqrt{\frac{s}{n}})$. It is sufficient to prove that for any $\epsilon > 0$, the smallest rate of $\xi_{n,p}$ is $\sqrt{\frac{s}{n}}$ such that we have:

$$P(\sup_{\|\boldsymbol{u}\|_2 = C} \bar{Q}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\beta}^1 + u\xi_{n,p}) < \bar{Q}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\beta}^1)) > 1 - \epsilon$$

where $\boldsymbol{u} \in \mathbb{R}^s$. This implies that with probability tending to 1, there is a local maximizer $\hat{\boldsymbol{\beta}}_1^{(1)}$ of the function $\bar{Q}$ in the neighborhood of $\boldsymbol{\beta}_0^1$ with the radius of $\boldsymbol{\beta}_0^1$ at most proportional to $\xi_{n,p}$.

$$\begin{aligned}
&\bar{Q}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\beta}_0^1 + u\xi_{n,p}) - \bar{Q}(\hat{\boldsymbol{\theta}}^{(0)}, \boldsymbol{\beta}_0^1) \\
=&-\tfrac{1}{2}\boldsymbol{u}^T\boldsymbol{X}^{1T}\dot{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}^{(0)})\boldsymbol{X}^1\boldsymbol{u}\xi_{n,p}^2 - (\mathbf{Z} - \boldsymbol{X}^1\boldsymbol{\beta}_0^1)\dot{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}^{(0)})\boldsymbol{X}^1\boldsymbol{u}\xi_{n,p} \\
&- np\sum_{j=1}^{m}\left(p'_{\lambda_{n,p}}(|\beta_{0j}|)sgn(\beta_j)u_{\beta_j}\xi_{n,p} + p''_{\lambda_{n,p}}(|\beta_{0j}|)u_{\beta_j}^2\xi_n^2)(1 + o(1))\right) \\
=&(I) + (II) + (III)
\end{aligned}$$

By Taylor's expansion, $\dot{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}^{(0)}) = \dot{\Sigma}^{-1}(\boldsymbol{\theta}_0) + \sum_{j=1}^{q}\dot{\Sigma}^j(\boldsymbol{\theta}*)$, where $\boldsymbol{\theta}*$ is a q dimension

vector between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}^{(0)}$. Therefore,

$$(I) = -\frac{1}{2}\boldsymbol{u}^T\boldsymbol{X}^{1T}\dot{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}^1\boldsymbol{u}\xi_{n,p}^2 - \frac{1}{2}\sum_{j=1}^{q}\boldsymbol{u}^T\boldsymbol{X}^{1T}\dot{\Sigma}^j(\boldsymbol{\theta}^*)\boldsymbol{X}^1\boldsymbol{u}\xi_{n,p}^2 u_j\eta_{n,p}$$

$$=(1)+(2)$$

and

$$(II) = -(\mathbf{Z}-\boldsymbol{X}^1\boldsymbol{\beta}_0^1)\dot{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}^{(0)})\boldsymbol{X}^1\boldsymbol{u}\xi_{n,p} + \sum_{j=1}^{q}(\mathbf{Z}-\boldsymbol{X}^1\boldsymbol{\beta}_0^1)\dot{\Sigma}^j(\boldsymbol{\theta}^*)\boldsymbol{X}^1\boldsymbol{u}\xi_{n,p}u_j\eta_{n,p}$$

$$=(3)+(4)$$

Noticing $\lambda_{\min}(\Sigma^{-1})>0$, $\sum_{i=1}^{n-1}\boldsymbol{X}_i^{1T}\boldsymbol{X}_i^1 = (\frac{n_1n_2}{n}-\frac{n_1^2}{n^2})I_s$ and

$$\left(\sum_{i=1}^{n-1}\boldsymbol{X}_{i=1}^{1T}\right)\left(\sum_{i=1}^{n-1}\boldsymbol{X}_{i=1}^{1}\right) = \frac{n_1^2}{n^2}I_s$$

we have

$$\begin{aligned}
(1) =& \frac{1}{2}\boldsymbol{u}^T\boldsymbol{X}^{1T}diag_{n-1}(\Sigma^{-1})(\tilde{I}_{n-1,p}+\tilde{J}_{n-1,p})\boldsymbol{X}^1\boldsymbol{u}\xi_{n,p}^2\\
=& -\frac{1}{2}\boldsymbol{u}^T\sum_{i=1}^{n-1}\boldsymbol{X}_i^{1T}\Sigma^{-1}\boldsymbol{X}_i^1\boldsymbol{u}\xi_{n,p}^2 - \frac{1}{2}\boldsymbol{u}^T\sum_{i=1}^{n-1}\boldsymbol{X}_i^{1T}\Sigma^{-1}\sum_{i=1}^{n-1}\boldsymbol{X}_i^1\boldsymbol{u}\xi_{n,p}^2\\
\leq& -\frac{1}{2}\boldsymbol{u}^T\left(\sum_{i=1}^{n-1}\boldsymbol{X}_i^{1T}\boldsymbol{X}_i^1 + \sum_{i=1}^{n-1}\boldsymbol{X}_i^{1T}\sum_{i=1}^{n-1}\boldsymbol{X}_i^1\right)\boldsymbol{u}\xi_{n,p}^2\lambda_{\min}(\Sigma^{-1})\\
=& -\frac{1}{2}\frac{n_1n_2}{n}\left\|\boldsymbol{u}\right\|_2^2\xi_{n,p}^2\lambda_{\min}(\Sigma^{-1})\\
\leq& -\frac{1}{2}\frac{\pi(1-\pi)}{\lambda_{\max}(\Sigma)}n\xi_{n,p}^2\left\|\boldsymbol{u}\right\|_2^2
\end{aligned}$$

By similar argument and A6(i), (2) $= O_p(n\xi_{n,p}^2\eta_{n,p}) = o_p((1))$ while $\eta_{n,p} = O_p(\sqrt{\frac{1}{np}})$.

For (II), by lemma 3.6.1 and A3,

$$\begin{aligned}(3) =& O_p(\sqrt{tr(\boldsymbol{X}^1\dot{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{X}^1)}\left\|\boldsymbol{u}\right\|_2\xi_{n,p})\\ =& O_p(\sqrt{\lambda_{\max}(\Sigma)\frac{n_1n_2}{n}tr(I_{s\times s})}\left\|\boldsymbol{u}\right\|_2\xi_{n,p})\\ =& O_p(\sqrt{ns}\left\|\boldsymbol{u}\right\|_2\xi_{n,p}))\end{aligned}$$

Similarly, (4) $= O_p(\sqrt{ns}\left\|\boldsymbol{u}\right\|_2\xi_{n,p}\eta_{n,p}) = o_p((3))$. So we have $(II) = O_p(\sqrt{ns}\left\|\boldsymbol{u}\right\|_2\xi_{n,p}))$.

$(III) = (5) + (6)$, where

$$(5) = -n\sum_{j=1}^{s} p'_{\lambda_{n,p}}(|\beta_{0j}|)sgn(\beta_j)u_{\beta_j}\xi_{n,p} \tag{3.6.41}$$

$$(6) = -n\sum_{j=1}^{s} p''_{\lambda_{n,p}}(|\beta_{0j}|)u_{\beta_j}^2\xi_n^2)(1+o(1))$$

Since $a_{n,p} = O_p(\frac{1}{\sqrt{n}})$ by A7,

$$|(5)| \le n\sqrt{s}a_{n,p}\left\|\boldsymbol{u}\right\|_2 = O_p(\sqrt{ns}\xi_{n,p}\left\|\boldsymbol{u}\right\|_2) \tag{3.6.42}$$

By A8

$$\begin{aligned}|(6)| &\le 2n\xi_{n,p}^2\sum_{j=1}^{s} p''(\beta_{0j})u_{\beta_j}^2 \le 2n\xi_{n,p}^2 b_{n,p}\left\|\boldsymbol{u}\right\|_2^2 \qquad (3.6.43)\\ &= o_p(n\xi_{n,p}^2)\end{aligned}$$

By choosing large $C = \|\boldsymbol{u}\|_2$, the smallest rate of ξ_{np} that (II) and (III) are dominated by (I) is $\xi_{n,p} = O_p(\sqrt{\frac{s}{n}})$. This completes the proof that $\left\|\hat{\boldsymbol{\beta}}_1^{(1)} - \boldsymbol{\beta}_0^1\right\|_2 = O_p(\sqrt{\frac{s}{n}})$.

step 2. in step 2 we prove that the vector $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^{(1)}, 0)$ is a strict local maximizer on $\mathbb{R}^d$. It is sufficient to prove for any given $\boldsymbol{\beta} \in \mathbb{R}^d$ satisfying $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 = O_p(\sqrt{\frac{s}{n}})$, we have $Q(\boldsymbol{\beta}^s, \hat{\boldsymbol{\theta}}^{(0)}) \geq Q(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}^{(0)})$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}^{1T}, \boldsymbol{\beta}^{2T})^T$ and $\boldsymbol{\beta}^s = (\boldsymbol{\beta}^{1T}, 0^T)^T$.

Let $\epsilon = C\sqrt{\frac{s}{n}}$, it is sufficient to prove for $j = s+1, s+2, ..., p$:

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}^{(0)})}{\partial \beta_j} &< 0 \ for \ 0 < \beta_j < \epsilon \\ \frac{\partial Q(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}^{(0)})}{\partial \beta_j} &> 0 \ for \ -\epsilon < \beta_j < 0 \end{aligned} \tag{3.6.44}$$

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}^{(0)})}{\partial \boldsymbol{\beta}_j} =& (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}_0)^T \dot{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}^{(0)}) \boldsymbol{X}_j + \sum_{l=1}^{p} \boldsymbol{X}_l^T \dot{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}^{(0)}) \boldsymbol{X}_j (\beta_l - \beta_{0l}) \\ & - nP'_\lambda(|\beta_j|) sgn(\beta_j) \\ =& (I) + (II) + (III) \end{aligned} \tag{3.6.45}$$

where $\boldsymbol{X}_j$ is the jth column of $\boldsymbol{X}$.

We first consider (I). By Taylor's expansion,

$$\begin{aligned} (I) =& (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}_0)^T \dot{\Sigma}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{X}_j + \sum_{k=1}^{q} (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}_0)^T \dot{\Sigma}^k(\boldsymbol{\theta}^*) \boldsymbol{X}_j u_k \eta_{n,p} \\ =& (5) + (6) \end{aligned}$$

Notice $\sum_{i=1}^{n-1} \boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij} = \frac{n_1 n_2}{n} - \frac{n_1^2}{n^2}$ and $(\sum_{i=1}^{n-1} \boldsymbol{X}_{ij})^T (\sum_{i=1}^{n-1} \boldsymbol{X}_{ij}) = \frac{n_1^2}{n^2}$.

$$
\begin{aligned}
(5) =& (\mathbf{Z} - \boldsymbol{X}\boldsymbol{\beta}_0)^T diag_{n-1}(\Sigma^{-1})(\tilde{I}_{n-1,p} + \tilde{J}_{n-1,p})\boldsymbol{X}_j = O_p(\sqrt{\boldsymbol{X}_j \dot{\Sigma}^{-1} \boldsymbol{X}_j}) \\
=& O_p(\sqrt{\sum_{i=1}^{n-1} \boldsymbol{X}_{ij}^T \Sigma^{-1} \boldsymbol{X}_{ij} + \sum_{i=1}^{n-1} \boldsymbol{X}_{ij}^T \Sigma^{-1} \sum_{i=1}^{n-1} \boldsymbol{X}_{ij}})
\end{aligned}
$$

where $\boldsymbol{X}_{ij}$ is the jth column of $\boldsymbol{X}_i$.

Noticing $\sum_{i=1}^{n-1} \boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij} = \frac{n_1 n_2}{n} - \frac{n_1^2}{n^2}$, $(\sum_{i=1}^{n-1} \boldsymbol{X}_{ij})^T (\sum_{i=1}^{n-1} \boldsymbol{X}_{ij}) = \frac{n_1^2}{n^2}$ and $\lambda_{\max}(\Sigma^{-1}) \leq \infty$, we have $(5) = O_p(\sqrt{n})$. Similarly, $(6) = O_p(\sqrt{n}\eta_{n,p}) = o_p((3))$, which is dominated by (5) if $\eta_{n,p} = o(1)$.

For (II), by Taylor's expansion,

$$
\begin{aligned}
(II) =& \sum_{l=1}^{p} \boldsymbol{X}_l^T \dot{\Sigma}^{-1}(\theta_0) \boldsymbol{X}_j (\beta_l - \beta_{0l}) + \sum_{k=1}^{q} \sum_{l=1}^{p} \boldsymbol{X}_l^T \dot{\Sigma}^k(\theta^*) \boldsymbol{X}_j (\beta_l - \beta_{0l})(\theta_k^* - \theta_{0k}) \\
=& (7) + (8)
\end{aligned}
$$

$$
\begin{aligned}
(7) =& \sum_{l=1}^{p} \boldsymbol{X}_l^T diag_{n-1}(\Sigma^{-1})(\tilde{I}_{n-1,p} + \tilde{J}_{n-1,p}) \boldsymbol{X}_j (\beta_l - \beta_{0l}) \\
=& \sum_{l=1}^{p} \left(\sum_{i=1}^{n-1} \boldsymbol{X}_{il}^T \Sigma^{-1} \boldsymbol{X}_{ij} + \sum_{i=1}^{n-1} \boldsymbol{X}_{il}^T \Sigma^{-1} \sum_{i=1}^{n-1} \boldsymbol{X}_{ij} \right) (\beta_l - \beta_{0l})
\end{aligned}
$$

Notice

$$
\begin{aligned}
& \sum_{i=1}^{n-1} \boldsymbol{X}_{il}^T \Sigma^{-1} \boldsymbol{X}_{ij} \leq \sum_{i=1}^{n-1} (\boldsymbol{X}_{il}^T \Sigma^{-1} \boldsymbol{X}_{il}))^{\frac{1}{2}} (\boldsymbol{X}_{ij}^T \Sigma^{-1} \boldsymbol{X}_{ij}))^{\frac{1}{2}} \\
\leq & \sum_{i=1}^{n-1} \lambda_{\max}(\Sigma^{-1}) (\boldsymbol{X}_{il}^T \boldsymbol{X}_{il}))^{\frac{1}{2}} (\boldsymbol{X}_{ij}^T \boldsymbol{X}_{ij}))^{\frac{1}{2}}
\end{aligned}
$$

$$=\lambda_{\max}(\Sigma^{-1})(\frac{n_1 n_2}{n}-\frac{n_1^2}{n^2})=O_p(n).$$

Also, let $B_l=\sum_{i=1}^{n-1}\boldsymbol{X}_{il}$. Then $B_l^TB_l=\left(\frac{n_1}{n}\right)^2$.

$$\sum_{i=1}^{n-1}\boldsymbol{X}_{il}^T\Sigma^{-1}\sum_{i=1}^{n-1}\boldsymbol{X}_{ij}=B_l^T\Sigma^{-1}B_j\leq(B_l^T\Sigma^{-1}B_l)^{1/2}(B_j^T\Sigma^{-1}B_j)^{1/2}\leq\lambda_{\max}(\Sigma^{-1})\frac{n_1^2}{n^2}$$

Then

$$(7)=O_p(n\left\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\right\|_2)=O_P(\sqrt{ns}) \tag{3.6.46}$$

Similarly, $(8)=O_p(\sqrt{ns}\eta_{n,p})$ Thus

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial\beta_j}=n\lambda_{n,p}\Big(O_P(\frac{\sqrt{s}}{\sqrt{n}\lambda_{n,p}}))+\frac{P'_{\lambda_{n,p}}(|\beta_j|)}{\lambda_{n,p}}sgn(\beta_j)\Big) \tag{3.6.47}$$

By assumption 9 and 10, the sign of 3.6.47 is determined by β_j, hence 3.6.44 followed. This implies $\hat{\boldsymbol{\beta}}^{(1)}$ should satisfy sparse property and completes the proof of step 2.

(d) Lastly we prove $\left\|\hat{\boldsymbol{\theta}}^{(1)}-\boldsymbol{\theta}_0\right\|_2=O_p(\sqrt{\frac{1}{np}})$. Since $\left\|\hat{\boldsymbol{\beta}}^{(1)}-\boldsymbol{\beta}_0\right\|_2=O_p(\sqrt{\frac{s}{n}})$, it is the same as the proof of (b). We omit the detail here and this completes the proof.

□

3.6.3 Proofs for consistency for PMLE with tappering

Lemma 3.6.3. *Assume A1, A2, A11 and A12 hold, we have:*

(1) $\left\|\Sigma(\boldsymbol{\theta})-\Sigma(\boldsymbol{\theta})_T\right\|_\infty=O(\frac{1}{p^{\delta_0}})$;

(2) $\left\|\Sigma_k(\boldsymbol{\theta}) - \Sigma_{k,T}(\boldsymbol{\theta})\right\|_\infty = O(\frac{1}{p^{\delta_0}})$;

(3) $\left\|\Sigma_{jk}(\boldsymbol{\theta}) - \Sigma_{jk,T}(\boldsymbol{\theta})\right\|_\infty = O(\frac{1}{p^{\delta_0}})$.

The matrix norm $\|\cdot\|_\infty$ *for the* $p \times p$ *matrix* $A = [a_{ij}]_{i,j=1}^p$ *is defined as the maximum of row sumation, i.e.* $\|A\|_\infty = \max_i \sum_{j=1}^p |a_{ij}|$

Proof. We show (1) in detail and omit the details for (2) and (3), as similar arguments can be applied.

$$\|\Sigma(\boldsymbol{\theta}) - \Sigma_T(\boldsymbol{\theta})\|_\infty = \max_i \sum_{j=1}^p \left|\gamma(h_{ij};\boldsymbol{\theta})K_T(h_{ij},w_p) - \gamma(h_{ij};\boldsymbol{\theta})\right| \tag{3.6.48}$$

where $h_{ij} = \left\|s_i - s_j\right\|_2$ is the distance between site s_i and s_j. For any $i = 1, 2, ..., p$,

$$\begin{aligned}
&\sum_{j=1}^p \left|\gamma(h_{ij};\boldsymbol{\theta})K_T(h_{ij},w_p) - \gamma(h_{ij};\boldsymbol{\theta})\right| \qquad (3.6.49)\\
&\leq \sum_{h_{ij}<w_p} \left|\gamma(h_{ij};\boldsymbol{\theta})K_T(h_{ij},w_p) - \gamma(h_{ij};\boldsymbol{\theta})\right| + \sum_{h_{ij}\geq w_p} \gamma_0(\boldsymbol{\theta},h_{ij})\\
&= (I) + (II)
\end{aligned}$$

Let $A^i = \{j : h_{ij} > w_p\}$ and $B_m^i = \{j : (m-1)\delta \leq h_{ij} < m\delta\}$, where Δ is independent of n and p. Then $A^i \subset \bigcup_{m=\lfloor \frac{w_p}{\Delta} \rfloor}^\infty B_m^i$. Let $V(R)$ be the volume of a d-dimensional baa of radius R. Then the volume of B_m^i is $B_m^i = V(m\delta) - V((m-1)\delta) = f_{d-1}(m)\delta^d$, where $f_{d-1}(m)$ is a polynomial function of m with degree of $d-1$. By A1, the number of sites in any unit subset of $D \subset \mathbb{R}^d$ is bounded, say ρ. Let $\#\{A\}$ denote the cardinality of a discrete set A. Then we have $\#\{B_m^i\} \leq f_{d-1}(m)\delta^d\rho$. Then exist a constant K such that

$f_{d-1}(m) \leq Km^{d-1}$ Then

$$
\begin{aligned}
(II) = \sum_{h_{ij} \geq w_p} \left|\gamma(\boldsymbol{\theta}, h_{ij})\right| &\leq \sum_{m=\lfloor \frac{w_p}{\delta} \rfloor}^{\infty} \sum_{j \in B_m^i} \left|\gamma(\boldsymbol{\theta}, h_{ij})\right| \\
&\leq K\rho \sum_{m=\lfloor \frac{w_p}{\delta} \rfloor}^{\infty} m^{d-1}\delta^d \max_{j \in B_m^i} \left|\gamma(\boldsymbol{\theta}, h_{ij})\right| \\
&\leq K\rho \int_{w_p}^{\infty} x^{d-1} \left|\gamma(\boldsymbol{\theta}, x)\right| dx \leq \frac{K\rho}{w_p} \int_0^{\infty} x^d \left|\gamma(\boldsymbol{\theta}, x)\right| dx
\end{aligned} \tag{3.6.50}
$$

Let $A_2^i = \{j : h_{ij} \leq w_p\}$. Then $A_2^i \subset \bigcup_{m=1}^{\lfloor \frac{w_p}{\delta} \rfloor + 1} B_m^i$.

$$
\begin{aligned}
(I) &= \sum_{h_{ij} < w_p} \left|\gamma(h_{ij}; \boldsymbol{\theta}) - \gamma(h_{ij}; \boldsymbol{\theta}) K_T(h_{ij}, w_p)\right| \\
&= 2 \sum_{h_{ij} < w_p} \left|\gamma(h_{ij}; \boldsymbol{\theta})\right| \frac{h_{ij}}{w_p} \\
&\leq \frac{2}{w_p} \sum_{m=1}^{\lfloor \frac{w_p}{\delta} \rfloor + 1} \sum_{j \in B_m^i} h_{ij} \left|\gamma(\boldsymbol{\theta}, h_{ij})\right| \\
&\leq \frac{2K\rho}{w_p} \sum_{m=1}^{\lfloor \frac{w_p}{\delta} \rfloor + 1} (m\delta)^{d-1}\delta \max_{j \in B_m^i} h_{ij} \left|\gamma(\boldsymbol{\theta}, h_{ij})\right| \\
&\leq \frac{2K\rho}{w_p} \int_0^{\infty} x^d \left|\gamma(\boldsymbol{\theta}, x)\right| dx \\
&\leq \frac{2K\rho}{w_p} \int_0^{\infty} x^d \gamma_0(\boldsymbol{\theta}, x) dx
\end{aligned} \tag{3.6.51}
$$

w_p has the same order as $p^{1/2}$ by A11. By A12, both (I) and $(II) = O(1/p^{1/2})$. This completes the proof. □

Lemma 3.6.4. *Assume A3-A6, A 11, A12 hold, we have*

(a) $\lim_{p\to\infty} \lambda_{\min}(\Sigma_T) > 0$, $\lim_{p\to\infty} \lambda_{\max}(\Sigma_T) < \infty$;

(b) There exists an open subset ω that contains the true parameter $\boldsymbol{\theta}_0$ such that for all $\boldsymbol{\theta}^* \in \omega$, we have:

(i) $-\infty < \lim_{p\to\infty} \lambda_{\min}(\Sigma_T^k(\boldsymbol{\theta}^*)) < \lim_{p\to\infty} \lambda_{\max}(\Sigma_T^k(\boldsymbol{\theta}^*)) < \infty$;

(ii) $-\infty < \lim_{p\to\infty} \lambda_{\min}(\Sigma_T^{kj}(\boldsymbol{\theta}^*)) < \lim_{p\to\infty} \lambda_{\max}(\Sigma_T^{kj}(\boldsymbol{\theta}^*)) < \infty$;

(iii) $\left|\frac{\partial t_{ij,T}(\boldsymbol{\theta}^*)}{\partial \theta_k}\right| = O(p)$ *for all* $k = 1, 2, ..., q$.

Proof. (a) Let $K_T = [K(h_{ij}, w)]_{i,j=1}^p$ be the tappering covariance. By eigenvalue inequalities of Schur product:

$$\min_{1\le i\le p} a_{ii}\lambda_{\min}(\Sigma) \le \lambda(\Sigma \circ K_T) \le \max_{1\le i\le p} a_{ii}\lambda_{\max}(\Sigma) \tag{3.6.52}$$

where a_{ij} are the $(i,j)th$ entry of matrix K_T. By A3, $\lambda_{\min}(\Sigma_T) > 0$ and $\lim_{p\to\infty} \lambda_{\max}(\Sigma_T) < \infty$

(b) Since $\Sigma_T^k = \Sigma^k \circ K_T$ and $\Sigma_T^{kj} = \Sigma^{kj} \circ K_T$ $[d](i)$ and $[d](ii)$ hold by A 6(ii). For [2](iii), since $t_{ij,T}(\boldsymbol{\theta}) = tr(\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T})$

$$\begin{aligned}\frac{\partial t_{ij,T}(\boldsymbol{\theta})}{\partial \theta_l} =& tr(\Sigma_T^{-1}\Sigma_{l,T}\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T}) + tr(\Sigma_T^{-1}\Sigma_{il,T}\Sigma_T^{-1}\Sigma_{j,T}) \\ &+ tr(\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{l,T}\Sigma_T^{-1}\Sigma_{j,T}) + tr(\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{jl,T}) \\ =& (1) + (2) + (3) + (4)\end{aligned} \tag{3.6.53}$$

Then (1) can be written as:

$$\begin{aligned}& tr(\Sigma_T^{-1}\Sigma_{l,T}\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T}) \\ =& tr((\Sigma_T^{-1} - \Sigma^{-1})\Sigma_{l,T}\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T}) + tr(\Sigma^{-1}(\Sigma_{l,T} - \Sigma_l)\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T}) +\end{aligned} \tag{3.6.54}$$

$$tr(\Sigma^{-1}\Sigma_l(\Sigma_T^{-1}-\Sigma^{-1})\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T}) + tr(\Sigma^{-1}\Sigma_l\Sigma^{-1}(\Sigma_{i,T}-\Sigma_i)\Sigma_T^{-1}\Sigma_{j,T}) +$$
$$tr(\Sigma^{-1}\Sigma_l\Sigma^{-1}\Sigma_i(\Sigma_T^{-1}-\Sigma^{-1})\Sigma_{j,T}) + tr(\Sigma^{-1}\Sigma_l\Sigma^{-1}\Sigma_i\Sigma^{-1}(\Sigma_{j,T}-\Sigma_j)) +$$
$$tr(\Sigma^{-1}\Sigma_l\Sigma^{-1}\Sigma_i\Sigma^{-1}\Sigma_j)$$

Define $\|\cdot\|_s$ for matrix A by $\|A\|_s = \max_i\{|\lambda_i(A)|, i = 1, 2, ..., p\}$, where $\lambda_i(A)$ is the ith eigenvalue of matrix A. Notice

$$\left\|\Sigma_T^{-1}-\Sigma^{-1}\right\|_s \le \left\|\Sigma^{-1}\right\|_s \|\Sigma-\Sigma_T\|_s \left\|\Sigma_T^{-1}\right\|_s .$$

Since $\lambda_{\min}(\Sigma^{-1}) = 1/\lambda_{\max}(\Sigma) > 0$, $\left\|\Sigma^{-1}\right\|_s \le \lambda_{\max}(\Sigma^{-1}) < \infty$. Also $\left\|\Sigma_T^{-1}\right\|_s < \infty$, $\left\|\Sigma_{j,T}\right\|_s < \infty$ for all $j = 1, 2, ..., q$. Hence $\left\|\Sigma_T^{-1}-\Sigma^{-1}\right\|_s = O_p(p^{-\delta}0)$. Then

$$\begin{aligned}
&\left|tr((\Sigma_T^{-1}-\Sigma^{-1})\Sigma_{l,T}\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T})\right| \\
&\le p\left\|((\Sigma_T^{-1}-\Sigma^{-1})\Sigma_{l,T}\Sigma_T^{-1}\Sigma_{i,T}\Sigma_T^{-1}\Sigma_{j,T})\right\|_s \\
&\le p\left\|(\Sigma_T^{-1}-\Sigma^{-1})\right\|_s \left\|\Sigma_{l,T}\right\|_s \left\|\Sigma_T^{-1}\right\|_s^2 \left\|\Sigma_{i,T}\right\|_s \left\|\Sigma_{j,T}\right\|_s \\
&= O(p/p^{\delta}0) = O(p^{1-\delta}0)
\end{aligned} \tag{3.6.55}$$

By similar argument, the first six terms in (1) all have the order of $O(p^{1-\delta}0)$. Apply the same argument on (2) − (4) we have:

$$\frac{\partial t_{ij,T}(\boldsymbol{\theta})}{\partial\theta_l} = \frac{\partial t_{ij}(\boldsymbol{\theta})}{\partial\theta_l} + O(p^{1-\delta}0) \tag{3.6.56}$$

By A6(iii), $\frac{\partial t_{ij,T}(\boldsymbol{\theta})}{\partial\theta_l} = O_p(p)$. This completes the proof.

□

Proof of Theorem 5. From lemma 4, all regularity conditions for Σ_T are satisfied. The proof of 5 is similar to that of Theorem 4. By replacing Σ by Σ_T and replacing A3-A6(*iii*) by the results in lemma 4, the results in Theorem 5 follows. □

3.6.4 Proofs for classification using PLDA

Lemma 3.6.5. *Let $\hat{\boldsymbol{\theta}}$ be the estimate of $\boldsymbol{\theta}_0$ and $\left\|\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_0\right\|_2 = O_p(\frac{1}{\sqrt{np}})$. Define*

$$\tilde{\Sigma} = \Sigma_T(\hat{\boldsymbol{\theta}}) = \Sigma(\hat{\boldsymbol{\theta}}) \circ K(w)$$

where $K(w)$ is defined in section 3.3.1.2. Assume A1, A2 and A11 and A12 hold,then

$$\left\|\tilde{\Sigma} - \Sigma\right\|_2 = O_p(c_n) \text{ and } \left\|\tilde{\Sigma}^{-1} - \Sigma^{-1}\right\|_2 = O_p(c_n)$$

where $c_{n,p} = max(\frac{w^d}{\sqrt{np}}, \frac{1}{w})$

Proof.

$$\begin{aligned}\left\|\tilde{\Sigma} - \Sigma\right\|_2 &= \left\|\Sigma(\hat{\boldsymbol{\theta}}) \circ K(w) - \Sigma(\boldsymbol{\theta}_0)\right\|_2 \\ &\leq \max_i \sum_{j=1}^{p} \left| r(\hat{\boldsymbol{\theta}}; h_{ij}) K_T(h_{ij}, w) - r(\boldsymbol{\theta}_0; h_{ij}) \right|\end{aligned}$$

where $K_T(h, w) = [(1 - h/w)_+]^2$. For any $i = 1, 2, ..., p$,

$$\begin{aligned}&\sum_{j=1}^{p} \left| r(\hat{\boldsymbol{\theta}}; h_{ij}) K_T(h_{ij}, w) - r(\boldsymbol{\theta}_0; h_{ij}) \right| \\ \leq &\sum_{h_{ij}<w} \left| r(\hat{\boldsymbol{\theta}}; h_{ij}) K_T(h_{ij}, w) - r(\boldsymbol{\theta}_0; h_{ij}) \right| + \sum_{h_{ij} \geq w} \left| r(\boldsymbol{\theta}_0, h_{ij}) \right|\end{aligned}$$

$$\leq \sum_{h_{ij}<w} \left|(r(\hat{\boldsymbol{\theta}}; h_{ij}) - r(\boldsymbol{\theta}_0; h_{ij}))K_T(h_{ij}, w)\right| + \sum_{h_{ij}<w} \left|r(\theta_0; h_{ij})K_T(h_{ij}, w) - r(\boldsymbol{\theta}_0; h_{ij})\right|$$

$$+ \sum_{h_{ij}\geq w} \left|r(\boldsymbol{\theta}_0, h_{ij})\right|$$

$$=(I) + (II) + (III)$$

From the same proof in lemma 3.6.3, we have $(II) = O_p(1/w)$ and $(III) = O_p(1/w)$. From A1 and A2(iii), we have

$$(I) \leq \sum_{h_{ij}<w} \left|r(\hat{\boldsymbol{\theta}}; h_{ij}) - r(\boldsymbol{\theta}_0; h_{ij})\right| \leq \sum_{k=1}^{q} \sum_{h_{ij}<w} \left|r_k(\boldsymbol{\theta}^*; h_{ij})\right| \left|\hat{\theta}_k^* - \theta_{0k}\right|$$

$$\leq Mw^d \rho \left\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_2$$

Therefore $(I) = O_p(\frac{w^d}{\sqrt{np}})$. Combine (I), (II) and (III), $\left\|\tilde{\Sigma} - \Sigma\right\|_2 = O_p(c_n)$.

Since $\Sigma(\boldsymbol{\theta}_0)$ and $\tilde{\Sigma} = \Sigma_T(\hat{\boldsymbol{\theta}})$ are positive definite, $\left\|\Sigma^{-1}\right\|_2 = \frac{1}{\lambda_{\min}(\Sigma)} < \infty$ and $\left\|\tilde{\Sigma}_T^{-1}\right\|_2 = \frac{1}{\lambda_{\min}(\tilde{\Sigma}_T)} < \infty$.

$$\left\|\tilde{\Sigma}_T^{-1} - \Sigma^{-1}\right\|_2 = \left\|\Sigma^{-1}(\Sigma - \tilde{\Sigma}_T)\tilde{\Sigma}_T^{-1}\right\|_2 \leq \left\|\Sigma^{-1}\right\|_2 \left\|\Sigma - \tilde{\Sigma}_T\right\|_2 \left\|\tilde{\Sigma}_T^{-1}\right\|_2 = O_p(c_n).$$

and this completes the proof. □

Lemma 3.6.6. *Assume A12 holds. Then* $\max_{1\leq i\leq s} \sum_{k=s+1}^{p} \sigma_{ik}^2$ *is bounded above by a constant.*

Proof. Since A12 holds, by similar argument as proving lemma 3.6.3, we have:

$$\begin{aligned}\max_{1\le i\le s}\sum_{k=s+1}^{p}\sigma_{ik}^2 &= \max_{1\le i\le s}\sum_{k=s+1}^{p}\gamma(h_{ik},\boldsymbol{\theta}_0) \le \max_{1\le i\le s}\left(\sum_{0<h_{ij}<1}\gamma(h_{ik},\boldsymbol{\theta}_0)+\sum_{h_{ij}\ge 1}\gamma(h_{ik},\boldsymbol{\theta}_0)\right)\\ &\le \int_0^1 h^{d-1}\gamma(h,\boldsymbol{\theta}_0)dh+\int_1^\infty h^{d-1}\gamma(h,\boldsymbol{\theta}_0)\\ &\le \int_{h=0}^1 h^{d-1}\gamma_0(h,\boldsymbol{\theta}_0)dh+\int_1^\infty h^{d}\gamma_0(h,\boldsymbol{\theta}_0)\le\infty\end{aligned}$$

□

Proof of Theorem 6. Suppose a new observation is from class 1, the conditional misclassification rate of $\hat{\delta}_{PMLE}$ for class 1 is:

$$W_1(\hat{\delta}_{PMLE};\Theta)=1-\Phi\left(\frac{(\boldsymbol{\mu}_1-\bar{\boldsymbol{Y}}-\frac{n_1-n_2}{2n}\hat{\boldsymbol{\Delta}})^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}}}{\sqrt{\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\Sigma\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}}}}\right) \tag{3.6.57}$$

Where $\bar{\boldsymbol{Y}}=\sum_{k=1}^{2}\sum_{i=1}^{n_k}\boldsymbol{Y}_{ki}$. We first consider denominator. From lemma 3.6.5, $\left\|\Sigma-\tilde{\Sigma}\right\|_2 = O_p(c_n)$ and $\left\|\Sigma^{-1}-\tilde{\Sigma}^{-1}\right\|_2 = O_p(c_n)$, where $c_n=\max(\frac{w^d}{\sqrt{np}},\frac{1}{w})$ and w is the threshold distance w. Then

$$\begin{aligned}\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\Sigma\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} =&\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}}+\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}(\Sigma-\tilde{\Sigma})\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} \\ \le&\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}}+\frac{\left\|\Sigma-\tilde{\Sigma}\right\|_2}{\lambda_{\min}(\tilde{\Sigma})}\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} \\ =&\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}}(1+O_p(c_n)) \\ =&(\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}+\hat{\boldsymbol{\Delta}}^T(\tilde{\Sigma}^{-1}-\Sigma^{-1})\hat{\boldsymbol{\Delta}})(1+O_p(c_n)) \\ =&\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}}(1+O_p(c_n))\end{aligned} \tag{3.6.58}$$

Write

$$\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}\Sigma^{-1}\boldsymbol{\Delta} + 2(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})^T\Sigma^{-1}\boldsymbol{\Delta} + (\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})^T\Sigma^{-1}(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}) \tag{3.6.59}$$

From Theorem 4, $\left\|\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}\right\|_2 = O_P(\sqrt{\frac{s}{n}})$. Hence $(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})^T\Sigma^{-1}(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta}) = O_P(\frac{s}{n})$. Also the second term

$$(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})^T\Sigma^{-1}\boldsymbol{\Delta} \le (\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta})^{\frac{1}{2}}\left((\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})^T\Sigma^{-1}(\hat{\boldsymbol{\Delta}}-\boldsymbol{\Delta})\right)^{\frac{1}{2}} \tag{3.6.60}$$

Since $\frac{s}{n\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}} \to 0$, we have

$$\begin{aligned}\hat{\boldsymbol{\Delta}}^T\Sigma^{-1}\hat{\boldsymbol{\Delta}} =& \boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1+O_P(\sqrt{\frac{s}{n\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}}}) + O_P(\frac{s}{n\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}})) \\ =& \boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1+O_P(\sqrt{\frac{s}{n\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}}}))\end{aligned} \tag{3.6.61}$$

Let $D_{n,p} = \max(\sqrt{\frac{s}{n\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}}}, c_n)$, the denominator can be represented by:

$$\sqrt{\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\Sigma\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}}} = \sqrt{\boldsymbol{\Delta}\Sigma^{-1}\boldsymbol{\Delta}(1+O_p(D_{n,p}))}.$$

Now consider the nominator.

$$\begin{aligned}&(\boldsymbol{\mu}_1 - \bar{\boldsymbol{Y}} - \frac{n_1-n_2}{2n}\hat{\boldsymbol{\Delta}})^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} \\ &= (\boldsymbol{\mu}_1 - \bar{\boldsymbol{Y}})^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} + \frac{n_2-n_1}{2n}\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} \\ &= (\boldsymbol{\mu}_1 - \bar{\boldsymbol{Y}} - \frac{n_2}{n}\boldsymbol{\Delta})^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} + \frac{n_2}{n}\boldsymbol{\Delta}^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} + \frac{n_2-n_1}{2n}\hat{\boldsymbol{\Delta}}^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}}\end{aligned} \tag{3.6.62}$$

$$= (1) + (2) + (3)$$

We start from (3). By lemma 3.6.5, $\hat{\boldsymbol{\Delta}}^T \tilde{\Sigma}^{-1} \hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\Delta}}^T \Sigma^{-1} \hat{\boldsymbol{\Delta}}(1 + O_p(c_n))$. From 3.6.61, (3) can be represented by $(3) = \frac{n_2 - n_1}{2n} \boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta}(1 + O_P(D_{n,p}))$.

For (2), first we have $\boldsymbol{\Delta}^T \tilde{\Sigma}^{-1} \hat{\boldsymbol{\Delta}} = \boldsymbol{\Delta}^T \Sigma^{-1} \hat{\boldsymbol{\Delta}}(1 + O_p(c_n))$. Then combine 3.6.61

$$\boldsymbol{\Delta}^T \Sigma^{-1} \hat{\boldsymbol{\Delta}} \leq \left(\boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta}\right)^{\frac{1}{2}} \left(\hat{\boldsymbol{\Delta}}^T \Sigma^{-1} \hat{\boldsymbol{\Delta}}\right)^{\frac{1}{2}} = \boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta}(1 + O_P(\sqrt{\frac{s}{n \boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta}}})) \quad (3.6.63)$$

Then (2) can be represented by: $(2) = \frac{n_2}{n} \boldsymbol{\Delta}^T \Sigma^{-1} \hat{\boldsymbol{\Delta}} = \frac{n_2}{n} \boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta}(1 + O_P(D_{n,p}))$.

Thus

$$(2) + (3) = \tfrac{1}{2} \boldsymbol{\Delta}^T \Sigma^{-1} \boldsymbol{\Delta}(1 + O_P(D_{n,p})) \quad (3.6.64)$$

Now we consider (1). Let $\hat{\boldsymbol{\Delta}} = (\hat{\boldsymbol{\Delta}}_1^T, \hat{\boldsymbol{\Delta}}_2^T)^T$ where $\hat{\boldsymbol{\Delta}}_1$ is s dimension and $\hat{\boldsymbol{\Delta}}_2$ is $p - s$ dimension. From Theorem 4, with probability tending to 1, $\hat{\boldsymbol{\Delta}}_2 = 0$ and $\left\| \hat{\boldsymbol{\Delta}} - \boldsymbol{\Delta} \right\|_2 = O_P(\frac{s}{n})$. Let $\xi = \boldsymbol{\mu}_1 - \bar{\boldsymbol{Y}} - \frac{n_2}{n} \boldsymbol{\Delta} = (\xi_1^T, \xi_0^T)^T$, where ξ_1 is s dimension and ξ_0 is $p - s$ dimension. Then $\xi \sim N(0, \frac{1}{n}\Sigma)$

Write

$$\Sigma = \begin{pmatrix} \Sigma_1 & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_2 \end{pmatrix}, \quad \Sigma^{-1} = \begin{pmatrix} C_1 & C_{12} \\ C_{12}^T & C_2 \end{pmatrix}$$

and

$$\tilde{\Sigma} = \begin{pmatrix} \tilde{\Sigma}_1 & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{12}^T & \tilde{\Sigma}_2 \end{pmatrix}, \quad \tilde{\Sigma}^{-1} = \begin{pmatrix} \tilde{C}_1 & \tilde{C}_{12} \\ \tilde{C}_{12}^T & \tilde{C}_2 \end{pmatrix}$$

where Σ_1, $\tilde{\Sigma}_1$, C_1 and $\tilde{C}_1$ are $s \times s$ matrix. Then

$$C_{12} = -\Sigma_1^{-1}\Sigma_{12}C_2 \text{ and } \tilde{C}_{12} = -\tilde{\Sigma}_1^{-1}\tilde{\Sigma}_{12}\tilde{C}_2.$$

Write

$$\begin{aligned}(1) =& \xi^T\tilde{\Sigma}^{-1}\hat{\mathbf{\Delta}} = \xi_1^T\tilde{C}_1\hat{\mathbf{\Delta}}_1 + \xi_0^T\tilde{C}_{12}\hat{\mathbf{\Delta}}_1 \\ =& \xi_1^T\tilde{C}_1\hat{\mathbf{\Delta}}_1 - \xi_0\tilde{C}_2\tilde{\Sigma}_2\tilde{\Sigma}_1^{-1}\hat{\mathbf{\Delta}}_1 \\ =& (i) + (ii)\end{aligned}$$

First we have: $\xi_1^T\tilde{C}_1\hat{\mathbf{\Delta}}_1 \leq (\xi_1^T\tilde{C}_1\xi_1)^{1/2}(\hat{\mathbf{\Delta}}_1^T\tilde{C}_1\hat{\mathbf{\Delta}}_1)^{1/2}$. By lemma3.6.5,

$$\xi_1^T\tilde{C}_1\xi_1 = \xi_1^T C_1\xi_1(1 + O_p(c_n))$$

Since $\xi_1 \sim N(0, \frac{1}{n}\Sigma_1)$, $E(\xi_1^T\Sigma_1^{-1}\xi_1) = tr(\frac{1}{n}I_s) = \frac{s}{n}$. Then $\xi_1^T\Sigma_1^{-1}\xi_1 = O_P(\frac{s}{n})$ and therefore $\xi_1^T C_1\xi_1^T \leq \xi_1^T\Sigma^{-1}\xi_1^T\frac{\lambda_{\max}(C_1)}{\lambda_{\min}(\Sigma_1^{-1})} = O_p(\sqrt{\frac{s}{n}})$. Hence $\xi_1^T\tilde{C}_1\xi_1 = O_P(\frac{s}{n})$.

Also,

$$\hat{\mathbf{\Delta}}_1^T\tilde{C}_1\hat{\mathbf{\Delta}}_1 = \hat{\mathbf{\Delta}}^T\tilde{\Sigma}^{-1}\hat{\mathbf{\Delta}} = \mathbf{\Delta}^T\Sigma^{-1}\mathbf{\Delta}(1 + O_p(D_{n,p}))$$

Hence (i) in (1) is: $(i) = \xi_1^T\tilde{C}_1\hat{\mathbf{\Delta}}_1 = (\mathbf{\Delta}^T\Sigma^{-1}\mathbf{\Delta})^{\frac{1}{2}}O_p(\sqrt{\frac{s}{n}})$.

The second term in (1) is:

$$(ii) = \xi_0\tilde{C}_2\tilde{\Sigma}_{12}\tilde{\Sigma}_1^{-1}\hat{\mathbf{\Delta}}_1 \leq (\hat{\mathbf{\Delta}}_1\tilde{\Sigma}_1^{-1}\hat{\mathbf{\Delta}}_1)^{1/2}(\xi_0\tilde{C}_2^T\tilde{\Sigma}_{12}^T\tilde{\Sigma}_1^{-1}\tilde{\Sigma}_{12}\tilde{C}_2)^{1/2}$$

By lemma 3.6.5

$$\xi_0^T \tilde{C}_2^T \tilde{\Sigma}_{12}^T \tilde{\Sigma}_1^{-1} \tilde{\Sigma}_{12} \tilde{C}_2 \xi_0 = \xi_0 C_2^T \Sigma_{12}^T \Sigma_1^{-1} \Sigma_{12} C_2 \xi_0 (1 + O_p(c_n))$$

Since $\xi_0 \sim N(0, \frac{1}{n}\Sigma_2)$,

$$\begin{aligned} E[\xi_0^T C_2^T \Sigma_{12}^T \Sigma_1^{-1} \Sigma_{12} C_2 \xi_0] \leq & \lambda_{\max}(\Sigma_1^{-1}) E[\xi_0^T C_2^T \Sigma_{12}^T \Sigma_{12} C_2 \xi_0] \\ = & \lambda_{\max}(\Sigma_1^{-1}) tr(E[\xi_0^T C_2^T \Sigma_{12}^T \Sigma_{12} C_2 \xi_0]) \\ = & \lambda_{\max}(\Sigma_1^{-1}) \frac{1}{n} tr(C_2^T \Sigma_{12}^T \Sigma_{12} C_2 \Sigma_2) \\ \leq & \lambda_{\max}(\Sigma_1^{-1}) \lambda_{\max}(\Sigma_2) \lambda_{\max}^2(C_2) \frac{1}{n} tr(\Sigma_{12} \Sigma_{12}^T) \end{aligned}$$

$$tr(\Sigma_{12}\Sigma_{12}^T) = \sum_{i=1}^{s} \sum_{k=s+1}^{p} \sigma_{ik}^2 \leq s \max_{1 \leq i \leq s} \sum_{k=s+1}^{p} \sigma_{ik}^2 \tag{3.6.65}$$

thus $\xi_0^T C_2^T \Sigma_{12}^T \Sigma_1^{-1} \Sigma_{12} C_2 \xi_0 = O_p(\frac{s}{n} d_{n,p})$, where $d_{n,p} = \max_{1 \leq i \leq s} \sum_{k=s+1}^{p} \sigma_{ik}^2$. By lemma 3.6.6, $d_{n,p}$ is bounded above by a constant. As a result:

$$\xi_0^T \tilde{C}_2^T \tilde{\Sigma}_{12}^T \tilde{\Sigma}_1^{-1} \tilde{\Sigma}_{12} \tilde{C}_2 \xi_0 = x i_0^T C_2^T \Sigma_{12}^T \Sigma_1^{-1} \Sigma_{12} C_2 \xi_0 = O_p(\frac{s}{n} d_{n,p})(1 + O_p(\max(c_n, \sqrt{\frac{s}{n}})))$$

Since

$$\hat{\mathbf{\Delta}}_1^T \tilde{\Sigma}_1^{-1} \hat{\mathbf{\Delta}}_1 \leq \hat{\mathbf{\Delta}}_1^T \tilde{C}_1 \hat{\mathbf{\Delta}}_1 \leq \hat{\mathbf{\Delta}}^T \tilde{\Sigma}^{-1} \hat{\mathbf{\Delta}} = \mathbf{\Delta}^T \Sigma^{-1} \mathbf{\Delta} (1 + O_P(D_{n,p})) \tag{3.6.66}$$

Let $A_{n,p} = \max(\sqrt{\frac{s}{n\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}}}, \frac{s}{n}, c_n)$, we have

$$(ii) = (\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta})^{\frac{1}{2}}O_P(A_{n,p})) \tag{3.6.67}$$

Combining the approximation of (i) and (ii) in (1), we have

$$(1) = \xi^T\tilde{\Sigma}^{-1}\hat{\boldsymbol{\Delta}} = (\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta})^{\frac{1}{2}}O_P(A_{n,p}).$$

As a result, since $\frac{\sqrt{s/n}}{Cp} \to 0$,

$$\begin{aligned} W_1(\hat{\delta}_{PLDA};\Theta) =& 1 - \Phi(\frac{\frac{1}{2}\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1+O_p(D_{n,p})) + \sqrt{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}}O_p(A_{n,p})}{\sqrt{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1+O_p(D_{n,p}))}}) \\ &= 1 - \Phi\left(\frac{\frac{1}{2}C_p(1+O_p(D_{n,p}) + \sqrt{C_p}O_p(A_{n,p})}{\sqrt{C_p}(1+O_p(D_{n,p}))}\right) \end{aligned} \tag{3.6.68}$$

Similarly, we can derive:

$$\begin{aligned} W_2(\hat{\delta}_{PLDA};\Theta) =& \Phi(\frac{-\frac{1}{2}\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1+O_p(D_{n,p})) + \sqrt{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}}O_p(A_{n,p})}{\sqrt{\boldsymbol{\Delta}^T\Sigma^{-1}\boldsymbol{\Delta}(1+O_p(D_{n,p}))}}) \\ &= \Phi\left(\frac{-\frac{1}{2}C_p(1+O_p(D_{n,p}) + \sqrt{C_p}O_p(A_{n,p})}{\sqrt{C_p}(1+O_p(D_{n,p}))}\right) \end{aligned} \tag{3.6.69}$$

Since $D_{n,p} \to 0$ and $A_{n,p} \to 0$, both $W_1(\hat{\delta}_{PLDA};\Theta)$ and $W_2(\hat{\delta}_{PLDA};\Theta)$ go to $1 - \Phi(\frac{\sqrt{C_0}}{2})$ as $n, p \to \infty$ with probability tending to 1 as $n, p, s \to \infty$. Then the approximate overall misclassification error rate is $W(\hat{\delta}_{PLDA};\Theta) = \frac{1}{2}(W_1(\hat{\delta}_{PLDA};\Theta) + W_2(\hat{\delta}_{PLDA};\Theta)) \to 1 -$

$\Phi(\frac{\sqrt{C_0}}{2})$. This completes the proof of sub-optimal.

Now we show the asymptotically optimal of $W(\hat{\delta}_{PLDA};\Theta)$. If $C_p \to C_0 < \infty$, $\frac{W(\hat{\delta}_{PLDA};\Theta)}{W_{OPT}} = \frac{W(\hat{\delta}_{PLDA};\Theta)}{\Phi(-\frac{\sqrt{C_p}}{2})} \to 1$.

If $C_p \to \infty$,

$$\frac{x\sqrt{C_p}}{4+x^2}e^{-\frac{x^2-C_p}{8}} \leq \frac{W(\hat{\delta}_{PLDA};\Theta)}{\Phi(-\frac{\sqrt{C_p}}{2})} \leq \frac{4+C_p}{x\sqrt{C_p}}e^{-\frac{x^2-C_p}{8}}$$

where $x = \frac{C_p(1+O_p(D_{n,p}))\pm 2\sqrt{C_p}O_p(A_{n,p})}{\sqrt{C_p(1+O_p(D_{n,p}))}} = \sqrt{C_p}(1+O_p(D_{n,p} \pm O(\frac{A_{n,p}}{\sqrt{C_p}})))$.

First $\frac{x\sqrt{C_p}}{4+x^2}e^{-\frac{x^2-C_p}{8}} \to 1$ and $\frac{4+C_p}{x\sqrt{C_p}} \to 1$ as $C_p \to \infty$. Also,

$$x^2 - C_p = C_p(O(D_{n,p} + O(\frac{A_{n,p}}{\sqrt{C_p}})))$$

if $C_p c_n \to 0$ and $C_p\sqrt{\frac{s}{n}} \to 0$, we have $x^2 - C_p \to 0$. Hence $\frac{W(\hat{\delta}_{PLDA};\Theta)}{W_{OPT}} \to 1$. This completes the proof. □

3.6.5 Remarks on the assumptions

Remarks on A3 : The first part of A3 is the same as that in [37]. We now verify that the cavariance matrix derived from *Matérn* covariance function satisfy the first part of A3. First, for symmetric matrix, we have

$$\lambda_{\max}(\Sigma) \leq (\|\Sigma\|_1)^{1/2}(\|\Sigma\|_\infty)^{1/2} = \|\Sigma\|_\infty = \max_i \sum_j^p \gamma(h_{ij})$$

Using the same notation in the proof of Lemma3.6.3, for each i

$$\sum_{j=1}^{p}\gamma(h_{ij}) \leq \sum_{m=0}^{\infty}\sum_{j\in B_m^i} r(h_{ij}) \leq K\rho\sum_{m=0}^{\infty} m^{d-1}\delta^d \max_{j\in B_m^i} r(h_{ij}) \leq K\rho\int_0^{\infty} h^{d-1}r(h)dh \tag{3.6.70}$$

Recall that *Matérn* covariance function has the following expansion at $h = 0$:

$$\gamma(h;\sigma^2,c,\nu,r) = \sigma^2(1-c)(1-b_1h^{2\nu}+b_2h^2+O(h^{2+2\nu}))\ as\ h\to 0$$

where b_1 and b_2 are explicit constants depending only on ν and r. Thus for $\epsilon > 0$,

$$\int_0^{\epsilon} h^{d-1}\gamma(h)dh = O(\int_0^{\epsilon} h^{d-1}dh) = O(\epsilon^d/d)\to 0\ as\ \epsilon\to 0 \tag{3.6.71}$$

Also, since $K_\nu(h)\propto e^{-h}h^{-\frac{1}{2}}(1+O(\frac{1}{h}))$ as $h\to\infty$, there exist a constant K, for any C sufficiently large, we have:

$$\int_C^{\infty} h^{d-1}\gamma(h)dh \leq K\int_0^{\infty} h^{d-1+v-\frac{1}{2}}e^{-h}dh = \Gamma(d+v-\tfrac{1}{2}) < \infty \tag{3.6.72}$$

3.6.71 and 3.6.72 lead to $\int_0^{\infty} h^{d-1}\gamma(h)dh < \infty$. Let $p\to\infty$ in 3.6.70, we have $\limsup_{p\to\infty}\lambda_{\max}(\Sigma) < \infty$ if Σ is derived from *Matérn* covariance function.

Now consider the second part of A3. A1 assumes increasing domain framework. Bachoc and Furrer (2016) [4] showed that under A1 and some weak assumptions on the matrix covariance function, if the spectral density of the covariance function is positive, the smallest eigenvalue of the covariance matrix is asymptotically bounded away from zero.

Most standard covariance function such as *Matérn* covariance function satisfy those assumptions hence satisfy the second part of A3.

Remarks on A4 and A5 : A4 and A5 are the same as the assumptions in [37]. $\|\Sigma_k\|_F = \sum_{i,j=1}^{p} \gamma_k^2(h_{ij};\boldsymbol{\theta})$, where $\gamma_k(h_{ij};\boldsymbol{\theta}) = \frac{\partial\gamma(h_{ij};\boldsymbol{\theta})}{\partial\theta_k}$, $k = 1, 2, ..., q$ and $\boldsymbol{\theta}$ is a k dimensional parameter. We now verify that *Matérn* covariance function satisfy A4 for fixed ν. For *Matérn* covariance function with fixed ν, we have

$$\begin{aligned}
\frac{\partial\gamma(h)}{\partial\sigma^2} &= \frac{2^{1-\nu}}{\Gamma(\nu)}(h/r)^\nu K_\nu(h/r)(1-c) \\
\frac{\partial\gamma(h)}{\partial c} &= -\sigma^2\frac{2^{1-\nu}}{\Gamma(\nu)}(h/r)^\nu K_\nu(h/r) \\
\frac{\partial\gamma(h)}{\partial(1/r)} &= \sigma^2(1-c)\frac{2^{1-\nu}}{\Gamma(\nu)}h(h/r)^\nu(2\frac{\nu}{h/r}K_\nu(h/r) - K_{\nu-1}(h/r))
\end{aligned} \tag{3.6.73}$$

It is easy to show that for each k, exist a constant $\epsilon > 0$ independent of n, p, for each i, there's j such that $\gamma_k(h_{ij}) > c$. As a result, $\|\Sigma_k\|_F = \sum_{i,j=1}^{p} \gamma_k^2(h_{ij};\boldsymbol{\theta}) \geq \sum_{i=1}^{p} \epsilon = p\epsilon$. Therefore $\|\Sigma_k\|_F^{-1} = O_p(p^{-1})$.

Remarks on A6 : We can also verify that the *Matérn* covariance function with fixed ν satisfy A6(i) and A6(ii). Similar to the procedure in the remarks on A3, it is sufficient to verify for any $\boldsymbol{\theta} \in \Theta$, $\gamma_k(h;\boldsymbol{\theta})$ and $\gamma_{kj}(h;\boldsymbol{\theta})$ belong to the function space:

$$\Im = \{f(x) : \int_0^\infty f(x)x^{d-1}dx < \infty\}$$

where $d \geq 1$ is the dimension of the domain. We have the first-order partial derivative of *Matérn* covariance function in (3.6.73). The second-order partial derivative of *Matérn*

covariance function is as follows:

$$
\begin{aligned}
&\frac{\partial^2\gamma(h)}{(\partial^2\sigma^2)} = 0 \\
&\frac{\partial^2\gamma(h)}{\partial\sigma^2\partial c} = -\frac{2^{1-\nu}}{\Gamma(\nu)}(h/r)^{\nu}K_{\nu}(h/r) \\
&\frac{\partial^2\gamma(h)}{\partial\sigma^2\partial(1/r)} = (1-c)\frac{2^{1-\nu}}{\Gamma(\nu)}h(h/r)^{\nu}(2\frac{\nu}{h/r}K_{\nu}(h/r) - K_{\nu-1}(h/r)) \\
&\frac{\partial^2\gamma(h)}{\partial c^2} = 0 \\
&\frac{\partial^2\gamma(h)}{\partial c\partial(1/r)} = -\sigma^2\frac{2^{1-\nu}}{\Gamma(\nu)}h(h/r)^{\nu}(2\frac{\nu}{h/r}K_{\nu}(h/r) - K_{\nu-1}(h/r)) \\
&\frac{\partial^2\gamma(h)}{\partial^2(1/r)} = \sigma^2(1-c)\frac{2^{1-\nu}}{\Gamma(\nu)}[(h/r)^{\nu-2}h^2K_{\nu}(h/r)(2\nu-1)2\nu \\
&\qquad - (4\nu+1)(h/r)^{\nu-1}h^2K_{\nu+1}(h/r) - (h/r)^{\nu}h^2K_{\nu+2}(h/r)]
\end{aligned}
\tag{3.6.74}
$$

Note that the covariance function and its first-order and second-order partial derivatives are linear combinations of a Bessel function of h times a polynomial of h. Similar to proving $\int_0^{\infty} h^{d-1}\gamma(h)dh < \infty$ in (3.6.70), we have $\gamma_k(h;\boldsymbol{\theta}) \in \Im$ and $\gamma_{kj}(h;\boldsymbol{\theta}) \in \Im$. Hence A6($i$) and A6($ii$) are satisfied. By similar procedure, we can verify that *Matérn* covariance function also satisfy A12 and A13.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Stanisław Adaszewski, Juergen Dukart, Ferath Kherif, Richard Frackowiak, Bogdan Draganski, et al. How early can we predict Alzheimer's disease using computational anatomy? *Neurobiology of aging*, 34(12):2815–2826, 2013.

[2] Yaman Aksu, David J Miller, George Kesidis, Don C Bigler, and Qing X Yang. An MRI-derived definition of MCI-to-AD conversion for long-term, automatic prognosis of MCI patients. *PLoS One*, 6(10):e25074, 2011.

[3] Juan Eloy Arco, Javier Ramírez, Juan Manuel Górriz, Carlos G Puntonet, and María Ruz. Short-term prediction of MCI to AD conversion based on longitudinal MRI analysis and neuropsychological tests. In *Innovation in Medicine and Healthcare 2015*, pages 385–394. Springer, 2016.

[4] François Bachoc and Reinhard Furrer. On the smallest eigenvalues of covariance matrices of multivariate spatial processes. *Stat*, 5(1):102–107, 2016.

[5] Peter J Bickel and Elizaveta Levina. Some theory for Fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010, 2004.

[6] Peter J Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

[7] Peter J Bickel and Elizaveta Levina. Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227, 2008.

[8] T. Tony Cai and Weidong Liu. Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, 106(494):672–684, 2011.

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

[10] T. Tony Cai, Zhao Ren, Harrison H Zhou, et al. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

[11] T. Tony Cai and Anru Zhang. Minimax rate-optimal estimation of high-dimensional covariance matrices with incomplete data. *Journal of multivariate analysis*, 150:55–74, 2016.

[12] T. Tony Cai, Cun-Hui Zhang, Harrison H Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

[13] T. Tony Cai and Linjun Zhang. High-dimensional linear discriminant analysis: Optimality, adaptive algorithm, and missing data1. *Technical report*, 2018.

[14] Tingjin Chu, Jun Zhu, Haonan Wang, et al. Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, 39(5):2607–2625, 2011.

[15] Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

[16] Robert W Cox. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

[17] Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

[18] Simon F Eskildsen, Pierrick Coupé, Daniel García-Lorenzo, Vladimir Fonov, Jens C Pruessner, D Louis Collins, et al. Prediction of Alzheimer's disease in subjects with mild cognitive impairment from the ADNI cohort using patterns of cortical thinning. *Neuroimage*, 65:511–521, 2013.

[19] Jianqing Fan and Yingying Fan. High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605, 2008.

[20] Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(4):745–771, 2012.

[21] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[22] Christine Fennema-Notestine, Donald J Hagler, Linda K McEvoy, Adam S Fleisher, Elaine H Wu, David S Karow, and Anders M Dale. Structural MRI biomarkers for preclinical and mild Alzheimer's disease. *Human brain mapping*, 30(10):3238–3253, 2009.

[23] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:, 2001.

[24] Peter Hall and Tapabrata Maiti. Choosing trajectory and data type when classifying functional data. *Biometrika*, page ass011, 2012.

[25] Miriam Hartig, Diana Truran-Sacrey, Sky Raptentsetsang, Alix Simonson, Adam Mezher, Norbert Schuff, and Michael Weiner. UCSF freesurfer methods. *ADNI: Alzheimers Disease Neuroimaging Initiative, San Francisco*, 2014.

[26] Roger Horn and Charles Johnson. Topics in matrix analysis. *ZAMM-Journal of Applied Mathematics and Mechanics*, 72(12):692–692, 1992.

[27] Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of magnetic resonance imaging*, 27(4):685–691, 2008.

[28] Clifford R Jack, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.

[29] David S Karow, Linda K McEvoy, Christine Fennema-Notestine, Donald J Hagler Jr, Robin G Jennings, James B Brewer, Carl K Hoh, and Anders M Dale. Relative capability of MR imaging and FDG PET to depict changes associated with prodromal and early Alzheimer's disease. *Radiology*, 256(3):932–942, 2010.

[30] Sang Han Lee, Alvin H Bachman, Donghyeon Yu, Johan Lim, Babak A Ardekani, et al. Predicting progression from mild cognitive impairment to Alzheimer's disease using longitudinal callosal atrophy. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 2:68–74, 2016.

[31] Kelvin K Leung, Josephine Barnes, Gerard R Ridgway, Jonathan W Bartlett, Matthew J Clarkson, Kate Macdonald, Norbert Schuff, Nick C Fox, Sebastien Ourselin, et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *Neuroimage*, 51(4):1345–1359, 2010.

[32] Kelvin K Leung, Jonathan W Bartlett, Josephine Barnes, Emily N Manning, Sebastien Ourselin, Nick C Fox, et al. Cerebral atrophy in mild cognitive impairment and Alzheimer disease rates and acceleration. *Neurology*, 80(7):648–654, 2013.

[33] Quefeng Li and Jun Shao. Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, pages 457–473, 2015.

[34] Yang Li, Yaping Wang, Guorong Wu, Feng Shi, Luping Zhou, Weili Lin, Dinggang Shen, et al. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. *Neurobiology of aging*, 33(2):427–e15, 2012.

[35] Qing Mai, Yi Yang, and Hui Zou. Multiclass sparse discriminant analysis. *arXiv preprint arXiv:1504.05845*, 2015.

[36] Qing Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, page asr066, 2012.

[37] Kanti V Mardia and Roger J Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.

[38] CR McDonald, LK McEvoy, L Gharapetian, C Fennema-Notestine, DJ Hagler, D Holland, A Koyama, JB Brewer, AM Dale, Alzheimers Disease Neuroimaging Initiative,

et al. Regional rates of neocortical atrophy from normal aging to early Alzheimer disease. *Neurology*, 73(6):457–465, 2009.

[39] C. D. Meyer. Matrix analysis and applied linear algebra siam, philadelphia, 2000. *Numerical Algorithms*, 26(2):198, 2001.

[40] Chandan Misra, Yong Fan, and Christos Davatzikos. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: results from ADNI. *Neuroimage*, 44(4):1415–1422, 2009.

[41] Hans-georg Müller. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics*, 32(2):223–240, 2005.

[42] Schabenberger Oliver and G Carol. Statistical methods for spatial data analysis, 2005.

[43] Rui Pan, Hansheng Wang, and Runze Li. Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association*, 111(513):169–179, 2016.

[44] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[45] Adam J Rothman, Peter J Bickel, Elizaveta Levina, Ji Zhu, et al. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

[46] Jun Shao, Yazhen Wang, Xinwei Deng, Sijian Wang, et al. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, 39(2):1241–1265, 2011.

[47] Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[48] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

[49] Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack, William Jagust, Enchi Liu, et al. The Alzheimer's Disease Neuroimaging Initiative: a review of papers published since its inception. *Alzheimer's & Dementia*, 9(5):e111–e194, 2013.

[50] Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.

[51] Daniela M Witten and Robert Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.

[52] Michael C Wu, Lingsong Zhang, Zhaoxi Wang, David C Christiani, and Xihong Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.

[53] Wei Biao Wu and Mohsen Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pages 1755–1768, 2009.

[54] Bradley T Wyman, Danielle J Harvey, Karen Crawford, Matt A Bernstein, Owen Carmichael, Patricia E Cole, Paul K Crane, Charles DeCarli, Nick C Fox, Jeffrey L Gunter, et al. Standardization of analysis sets for reporting results from ADNI MRI data. *Alzheimer's & dementia: the journal of the Alzheimer's Association*, 9(3):332–337, 2013.

[55] Peirong Xu, Ji Zhu, Lixing Zhu, and Yi Li. Covariance-enhanced discriminant analysis. *Biometrika*, 102(1):33–45, 2014.

[56] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

[57] Daoqiang Zhang, Dinggang Shen, et al. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PloS one*, 7(3):e33182, 2012.

[58] Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of statistics*, 36(4):1509, 2008.