Some Advances in Classifying and Modeling Complex Data

Angang Zhang

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Xinwei Deng, Chair
Yili Hong
Eric Smith
Inyoung Kim

December 12, 2015
Blacksburg, Virginia

This thesis is dedicated to my parents.

Acknowledgements

Dedicated to my parents, Zhang Kuan and Lin Hua for their endless support over my graduate school years.

I would like to express my deepest gratitude to my advisor and committee chair, Dr. Xinwei Deng, who has guided me through my research work and conveyed the spirit of the way to think deeply. Without his guidance and persistent help, this dissertation would not have been possible. Within the three years of academic cooperation, he patiently led me through different areas of statistical research, ways to overcome difficulty as well as writing and presenting skills. I'm very grateful to have the chance to working with him.

I would like to thank my committee members, Dr. Yili Hong, Dr. Inyoung Kim and Dr. Eric P. Smith. Their advices and supportive suggestions have broaden my view of certain subjects and made my dissertation writing a little less amateurish. I would like to thank the statistics department of Virginia Tech for the systematic training of my statistics skills. I would also like to thank Dr. Eric A. Vance for providing me the opportunity of statistics consulting in the laboratory for interdisciplinary statistical analysis (LISA). His advice and help made my consultant experience an enjoyable one.

In addition, a thank you to Justin Hobart and Justin Wang and the whole decision scientist group of Microsoft, who supported me patiently on the two stage risk model building and evaluation project. I really appreciate that they provided me with wonderful guidance and gave me permission to use the online purchase data in my dissertation.

I would like to thank my family, in particular my parents. It is your unconditional love and constant faith in me that raised me up and bring me this far. I would like to thank them for giving me the strength to overcome all difficulties in my life and my work. Thanks for supporting me for my studies and all their understanding and encouragements along the five years despite the long distance between us. It is their love that made me realize this achievement.

To all my friends, thank them for the encouragements along the way. I cannot list all the names here but their friendship will always be on my mind. Special thanks to my girlfriend, Jingjing Liu, who has been most understanding and tolerating for the past year. It is my honour to have she accompanying me along the way.

Abstract

In statistical methodology of analyzing data, two of the most commonly used techniques are classification and regression modeling. As scientific technology progresses rapidly, complex data often occurs and requires novel classification and regression modeling methodologies according to the data structure. In this dissertation, I mainly focus on developing a few approaches for analyzing the data with complex structures.

Classification problems commonly occur in many areas such as biomedical, marketing, sociology and image recognition. Among various classification methods, linear classifiers have been widely used because of computational advantages, ease of implementation and interpretation compared with non-linear classifiers. Specifically, linear discriminant analysis (LDA) is one of the most important methods in the family of linear classifiers. For high dimensional data with number of variables p larger than the number of observations n occurs more frequently, it calls for advanced classification techniques. In Chapter 2, I proposed a novel sparse LDA method which generalizes LDA through a regularized approach for the two-class classification problem. The proposed method can obtain an accurate classification accuracy with attractive computation, which is suitable for high dimensional data with $p > n$.

In Chapter 3, I deal with the classification when the data complexity lies in the non-random missing responses in the training data set. Appropriate classification method needs to be developed accordingly. Specifically, I considered the "reject inference problem" for the application of fraud detection for online business. For online business, to prevent fraud transactions, suspicious transactions are rejected with unknown fraud status, yielding a training data with selective missing response. A two-stage modeling approach using logistic regression is proposed to enhance the efficiency and accuracy of fraud detection.

Besides the classification problem, data from designed experiments in scientific areas often have complex structures. Many experiments are conducted with multiple variance sources. To increase the accuracy of the statistical modeling, the model need to be able to accommodate more than one error terms. In Chapter 4, I propose a variance component mixed model for a nano material experiment data to address the between group, within group and within subject variance components into a single model. To adjust possible systematic error introduced during the experiment, adjustment terms can be added. Specifically a group adaptive forward

and backward selection (GFoBa) procedure is designed to select the significant adjustment terms.

Key Words: A/B testing, fraud detection, linear classifier, misclassification error, net profit value, reject inference, sparse linear discriminant analysis, two-class classification, variance component mixed model.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Statistical methodology are widely used for analyzing the data and helping draw proper conclusions. Perhaps the two most commonly used techniques are classification and regression modeling. Classification (Ethem, 2004) is a problem of identifying the group label for observations with unknown class labels. For data with general structure, regression modeling techniques are usually employed for estimation and prediction. As scientific technology advances, complex data occurs in many areas. Three types of complex data are discussed in this dissertation, including high dimensional data with the number of variables p larger than the number of observations n, a data set with non-random missing class labels, and data from designed experiment with multiple sources of variation. Those complex data usually call for advanced classification and modeling methodologies that are designed for the data structure.

For classification problems, data is collected with class labels indicating which category each observation belongs to. The classification rule can be used for identifying the classes for data with unknown class labels. Based on the context and background of different classification problem, various classification methodologies have been developed. A classification method is referred to as a classifier. Based on the form of the classifier, classification methodologies can generally be grouped into linear classifiers and non-linear classifiers, The linear classifiers employ linear combination of the independent variables for constructing the classification rule.

For conducting classification, it is comparatively straightforward if there are no missing responses in the training data set. While in some applications such as the credit card application problem (Bolton and Hand, 2002; Delamaire et al., 2009), the training data contains observations that have unknown class labels. Under this context, the traditional classifiers may not be appropriate, especially when the missing class labels are not randomly missing. novel statistical techniques need to be developed in the situation of training data with both labelled and unlabelled observations.

For general data with the continuous responses, there is a certain type of data collected from carefully designed and controlled experiments. There are often multiple sources of variance introduced during the experimental procedure. To get accurate analysis and draw valid conclusions, it often needs meticulous modeling techniques to uncover the variability of the data with respect to multiple components. By taking into account the mixed variance structure, I will develop a mixed variance component model to analyze the data with multiple sources of variances. The proposed model is applied to data from a nano experiment from material engineering.

In this chapter, I will first review several existing methodologies of the linear classifiers in Section 1.1. In Section 1.2, classification with training data containing missing class labels are introduced in the content of reject inference. Section 1.3 introduces some modeling techniques for data with multiple sources of variance.

1.1 Linear Classifiers

Classification has broad applications in many areas. In the biomedical and bioinformatics areas, statistical methods are used for finding the relative genes that controls certain characteristics (Liu and Yu, 2005; Au et al., 2005). To improve customer service, classification methodologies are used for customer relationship management (CRM) (Ngai et al., 2009). In sociology studies, classification is used to study characteristics of different social classes (Rose and Pevalin, 2001). Other application areas of classification includes public health, engineering and image recognition.

Linear classifiers are an important category in the classification methodologies. They usually enjoy the advantages of fast computation, ease of implementation and interpretation. Suppose there are k classes in total, p independent variables and n observations. Denote by $X_1, \cdots, X_p$ the independent variables and Y the class label. A linear classifier predicts the classes by $p_k = f_k(\sum_i \beta_i X_i)$ such that a linear combination of the independent variables is used, where f_k is a function that maps a linear combination of X_i to the probability of p_k that

$\boldsymbol{x} = (X_1, \cdots, X_p)'$ belongs to class k.

One of the simplest linear classifier, resulted from the Fisher discriminant analysis (FDA) (introduced by R. A. Fisher (1936)) searches for a linear separating hyperplane $\boldsymbol{\beta}$ that maximizes $\frac{\boldsymbol{\beta}'\boldsymbol{\Sigma}_b\boldsymbol{\beta}}{\boldsymbol{\beta}'\boldsymbol{\Sigma}_w\boldsymbol{\beta}}$, where $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_w$ are pooled within and between class sample variance covariance matrices, respectively. Equivalently, under the assumption that observations in each class follow multivariate normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ with a same covariance structure $\boldsymbol{\Sigma}$ for different classes, Linear Discriminant Analysis (McLachlan, 2004; Hastie et al., 2008) is proposed that forms a linear decision boundary.

Besides LDA, another widely used linear classifier is logistic regression (Hosmer and Lemeshow, 1989; Hastie et al., 2008). It models the posterior probability for an observation $\boldsymbol{x}$ to belong to class k as

$$P(Y = k|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\beta}_k'\boldsymbol{x})}{1 + \sum_{i=1}^{K-1} \exp(\boldsymbol{\beta}_i'\boldsymbol{x})}, k = 1, \cdots, K-1,$$
$$P(Y = K|\boldsymbol{x}) = \frac{1}{1 + \sum_{i=1}^{K-1} \exp(\boldsymbol{\beta}_k'\boldsymbol{x})},$$

where functions of linear combinations of independent variables $\boldsymbol{x}$ are defined to construct the classifier. As a special case of the logistic regression, the multinomial naive Bayes classifier (Ng and Jordan, 2002) assumes that the independent variables are conditionally independent given the class information. Under the multinomial assumption, the naive Bayes classifier becomes a linear classifier with its log likelihood function proportional to a linear combination of independent variables (McCallum and Nigam, 1998). Other extensions of logistic regression include ordered logistic regression (Hardin and Hilbe, 2007), and mixed logit model (McFadden and Train, 2000).

A special case of the neural network (Auer et al., 2008), a single-layer perceptron (Gallant, 1990; Freund and Schapire, 1999) also constructs a linear classifier. It assigns weight to each independent variable and maps the feature space to the class label by $\hat{y} = \arg\max_k f(\boldsymbol{x}, k)\boldsymbol{w}$,

where $f(\boldsymbol{x}, k)$ is a linear function of $\boldsymbol{x}$ and $\boldsymbol{w}$ is a weight vector. Extended by modifying the algorithm of perceptron, the pocket algorithm (Gallant, 1990) chooses the best classifier in each iteration rather than the final classifier until the convergence of each iteration. While Winnow (Nick, 1988), another linear classifier extended from the perceptron algorithm, updates the weight $\boldsymbol{w}$ in a multiplicative way.

Another commonly used linear classifier, the linear support vector machine (SVM) (Cortes and Vapnik, 1995; Hsu and Lin, 2002) constructs a linear classifier by solving the optimization function:

$$\min ||\boldsymbol{w}||, \text{ subject to } y_i(\boldsymbol{x}_i'\boldsymbol{w} - b) \geq 1, i = 1, \cdots, n,$$

where $\boldsymbol{w}$ is the weight vector that makes the classifier a linear classifier. Extensions of SVM includes transductive SVM (Joachims, 1999) and structured SVM (Finley and Joachims, 2008).

As data collection techniques advance, it is often observed that the number of independent variables p exceeds the number of observations n in the so-called high dimensional data problems. There are different linear classifiers proposed under the high dimensional data context since linear classifiers enjoy advantages of faster calculation speed and simpler form of classification rule. One of the simplest linear classifier, LDA has been extended widely for high dimensional data. The nearest shrunken centroids (NSC) method proposed by Tibshirani et al. (2003) performs regularization on the mean structure of each class. Based on the soft thresholding idea, NSC uses the shrunken centroids vector to substitute for the sample mean $\boldsymbol{\mu}_k$ and reduces the number of predictors. NSC has been widely implemented in genomic data and microarray problems (Dabney, 2005; Wang et al, 2007). Several extensions of NSC have been proposed including shrunken centroids regularized discriminant analysis (SCRDA) (Guo et al., 2007), l_∞-norm penalized NSC (ALP-NSC) and the adaptive hierarchical penalized NSC (AHP-NSC) (Wang and Zhu, 2007). Besides penalizing on mean structure, there are multiple methods that propose penalizing on the covariance structure. In Witten and Tibshirani (2009), Scout is proposed with a penalizing term on $\boldsymbol{\Sigma}^{-1}$. In Shao et al. (2011), penalizing directly on $\boldsymbol{\Sigma}$ is proposed. Other methods that apply penalization terms on the

covariance structure includes Guo et al. (2007), Rothman et al. (2008).

There are classification methods that are proposed under other contexts. A penalized method based on Fisher's discriminant analysis can be found in Witten and Tibshirani (2011). Sparse LDA (sLDA) (Wu et al., 2009) estimates the discriminant direction by maximizing the Rayleigh quotient with certain l_1 and l_2 constraints. Mai et al. (2012) proposed a classification rule by obtaining a penalized least square estimate for the classification direction. It has been revealed in Mai and Zou (2013) for the equivalence of sLDA (Wu et al., 2009) with the sparse discriminant analysis in Clemmensen et al. (2011) and Mai et al. (2012). A median based classifier which is expected to have robust performance is proposed in Hall et al. (2009).

In Chapter 2, I proposed a classification method extending from LDA for high dimensional data. The proposed method is developed specifically for two-class classification. It can maintain the sparse structure of the inverse covariance matrix $\mathbf{\Sigma}^{-1}$ and the mean difference between the two classes simultaneously. More of the related background and statistical methodologies for the two-class classification problem will be introduced there.

1.2 Reject Inference

Under the conventional setting of classification, there is no missing data in the training data set. That is, the observations in the training data contains complete label information. Specifically, there is no missing information on the class labels. There are circumstances where the training data contains observations with unknown class labels. They are included either for the purpose of increasing the classification accuracy by making use of the unlabeled data or there is difficulty in collecting the label information for all observations. Some applications include human behaviour learning (Zhu et al., 2007) and image learning (Camps-Valls et al., 2004; Guillaumin et al., 2010).

The main focus of this chapter is to handle data with non-random missing class labels usually referred to as reject inference (Chen and Astebro, 2001; Feelders, 2003; Crook and Banasik, 2004). It originates from a real practical problem: fraud detection. The training

data consists of transactions that occur during online commercial activities. Transactions suspected to be fraud are rejected such that the corresponding normal/fraud class labels are missing, yielding a training data set with selective missing class labels. Usually a risk model is constructed as the means of fraud detection.

Reject inference has been widely studied in different applications. Under the context of online business, different reject inference methods can be generally grouped into three categories in terms of model assumptions. The first category assumes that the accepted transactions can represent the whole population of all transactions and contain the full information for risk modeling. This assumption is very strong and often too expensive to be satisfied. The A/B testing (Kohavi et al., 2012), is a known technique of this kind. If properly used, A/B testing can lead to an accurate risk model. However, this technique is costly since all the fraud cases in group B are accepted during the A/B testing procedure.

The other two types of model assumptions are referred to as missing at random (MAR) and missing not at random (MNAR) (Feelders, 2003), respectively. A thorough discussion of the missing schemes in reject inference can be found in Shaun et. al. (2013). Under MAR, the fraud status of rejected transaction is missing at random (Feelders, 2003). For each transaction, denote by a its acceptance status with value 1 being accepted, and 0 otherwise. Similarly, define y as its fraud status of whether this transaction is fraud or not (labelled as 1 or 0, respectively). Then the assumption of MAR can be written as

$$P(y = 1|\boldsymbol{x}, a = 1) = P(y = 1|\boldsymbol{x}, a = 0),$$

where $\boldsymbol{x}$ is a vector of predictor variables of the transaction. In another words, the MAR model assumes that the rejected samples and the accepted samples come from the same population.

There are several methods proposed under the assumption of MAR. One commonly used method is called augmentation (Montrichard, 2008; Banasik and Crook, 2007). The corresponding risk model is constructed by reweighing the accepted transactions. For example, Hsia (1978) assumed that the percentage of fraud among the accepted transactions is the same

as that among the rejected transactions. This assumption is unrealistic in practice. Another augmentation method (Montrichard, 2008) employs a two-step modeling process. It first models the acceptance/rejection status for all the transactions. Then a weight of $1/P(a = 1)$ is imposed on the accepted transactions for classification of being fraud or not. This method can result in large bias in model estimation due to large weight $1/P(a = 1)$ on some observations.

Another widely used approach under MAR is called extrapolation (Hand and Henley, 1993). In Siddiqi (2006), a classification model is constructed to classify the normal/fraud status by only using the accepted observations. Then the rejected observations are scored according to this model. Feelders (1999) assumed that the distributions of risk score for the rejected observations and accepted observations came from a particular family of distribution. Then the unknown fraud/non-fraud status for rejected observations are treated as missing data. However, the assumption behind the method is questionable when the accepted and rejected populations are not homogeneous.

Note that the assumption of MAR often does not hold in reality because of a selective reject nature. A more reasonable assumption is missing not at random (MNAR). The MNAR assumes that the distribution of rejected population is different from that of the accepted population:

$$P(y = 1|\boldsymbol{x}, a = 1) \neq P(y = 1|\boldsymbol{x}, a = 0).$$

Joanes and Derrick (1993) developed an iterative method under this framework. They constructed a prior probability of being fraud in the rejection region. Then they iteratively conduct the classification and modify the prior probability. Another example under MNAR is called two-stage bivariate probit model (Heckman, 1979; Copas and Li, 1997; Chen, 2001). This method separates the fraud/non-fraud status and the good/bad status into two models and links them together under the bivariate normal error assumption:

$$a = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon_1$$

$$y = \boldsymbol{x}'\boldsymbol{\gamma} + \epsilon_2,$$

Note that the fraud status y is only observed when the acceptance status $a = 1$. The error terms ϵ_1 and ϵ_2 follow a bivariate normal distribution. However, the estimation in the bivariate probit model is not very robust and the multicollinearity problem often exists (Puhani, 2000).

I propose a two-stage method to build the risk model for reject inference under the assumption of MNAR. The proposed two-stage model differs from Heckman′s two-stage bivariate probit model in that the acceptance/rejection and the fraud/non-fraud information are used in both stages. Certain background for online business fraud detection and the proposed two-stage model will be detailed in Chapter 3.

1.3 Mixed Variance Component Model

Various scientific fields need experiments to be carefully designed and conducted. Statistical analysis has been used as an important tool for interpreting the results of the experiments. Many experiments have been designed to include multiple measurements on each subject and multiple replicates under each treatment level such that multiple variance sources are included naturally. To address the multiple variance components, variance component mixed model can be constructed to include more than one error term (Robinson, 1987).

A typical example for the use of variance component mixed model lies in the longitudinal study in drug test experiments (Stram and Lee, 1994; Fitzmaurice et al., 2004). Drug test experiments are usually conducted on multiple subjects with some subjects supplied with drug and the others with placebo. The drug effect is then recorded over time via certain indicative indices. To address the drug effect over time as well as the random subject effect, usually two variance components can be included in a mixed effects regression model (Laird and Ware 1982):

$$Var(\boldsymbol{y}_i|\boldsymbol{X}_i) = \boldsymbol{Z}_i\boldsymbol{\Sigma}_v\boldsymbol{Z}_i' + \sigma_\epsilon^2\boldsymbol{\Omega}_i,$$

where $\boldsymbol{Z}_i$ is the design matrix, $\boldsymbol{\Sigma}_v$ is the random subject variance component, σ_ϵ^2 controls the size of the time effect variance component, $\boldsymbol{\Omega}_i$ is the time effect variance component, $\boldsymbol{y}_i$ is the vector of indicative index measured over time for subject i and $\boldsymbol{X}_i$ is the independent

characteristic of the ith subject. For measurements on a same subject over time, it is normal to assume that the covariance between two points is decreasing when the time lag is further away. Covariance structures that are commonly used for $\mathbf{\Omega}_i$ includes independent error structure, AR(1), or Toeplitz (Littell et al., 2000). The variance component mixed model has been applied widely to medical studies such as pharmaceutical science (Bonate, 2006), disease detection (Gao et al., 1998) and epidemiologic (Takkouche et al., 1999).

Experiments emerge that have a similar or more complicated error structure in other areas. Mixed variance component model has been developed for the specific experiment. In biological study, the mixed model is used for identifying the variance introduced by different gene sets (Visscher et al., 2006; Meyer, 1999; George et al., 2000). In ecology, different landforms have effect on the animal population density which can be addressed by including another variance component (Bolker et al., 2007; Nussey et al., 2007). In chemistry, experiments are performed on chemical substance considering individual variation (Hrushka et al., 2005). In sociology area, where human behaviours are frequently studied, subject variation can be included as another variance component (Kellam et al., 2008). Mixed variance component models have also been used in engineering. For example, multiple error sources are included in the design of orbit determination algorithms (Tapley et al., 2004).

Nanomaterials possess great mechanical properties with wide applications in many areas (Yu et al., 2000). Experiments are often conducted for measuring certain mechanical properties of interest (Bogner et al., 2006). How to accurately quantify mechanical properties of nanomaterials is thus very important but challenging due to nanoscale manipulation and tactful measurement techniques. Statistical modeling approach combining physical theories have been used for the quantification of nanomaterials (Deng et al., 2009; Mai and Deng, 2010). In Chapter 4, a nano material experiment is introduced with different force levels as experimental treatment groups. Under each treatment group, multiple profiles will be generated as multiple replications. Thus data from the nano material experiment contains three variance components: the between group variance component, the within group variance component and the within profile variance component. I propose a novel mixed variance component

model to accommodate experimental variations and artifacts for analyzing the nanomaterial experiment data. The proposed method can automatically adjust systematic errors in the experiments through a group adaptive forward backward selection (GFoBa). It thus leads to accurate estimation of mechanical properties with the ability to filter out various experimental errors. The performance of the proposed methods is compared with other existing method through both simulation and a real data example. Certain backgrounds of the experiment and the algorithm for the mixed variance component model will be detailed in Chapter 4.

Chapter 2 A Regularized Approach to Sparse Linear Discriminant Analysis for Two-class Classification

2.1 Introduction

Classification with two classes widely occur in many areas (Ethem, 2004; Hastie, Tibshirani and Friedman, 2008) such as handwriting recognition, credit scoring, biological classification and disease detection. Among various two-class classification methods, discriminant analysis is one of most popular methods used in practice because of good prediction performance with easy implementation and interpretation (Morrison, 1969; Lachenbruch, 1979). Based on the Bayes classification rule, the traditional linear discriminant analysis (LDA) forms a classifier with linear decision boundary (Fisher, 1936). Denote the variables by $\boldsymbol{x}$ and the two classes by Y, $Y = 1, 2$, throughout the chapter. Under the assumption that the conditional density function $f(\boldsymbol{x}|Y = k)$ follows normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, $k = 1, 2$, the LDA classification rule can be obtained by

$$\arg\max_k \left[\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k \right], \tag{2.1}$$

where π_k is the prior probability for a sample class k.

The LDA can have good performance for the situation where the number of covariates p is smaller than the sample size n. It has been shown that LDA enjoys asymptotic properties that as n goes to infinity, the misclassification rate goes to zero (Bickel and Levina, 2008; Fan and Fan, 2008; Shao et al., 2011). As scientific technology progresses rapidly, high dimensional data with thousands of covariates are often encountered in practice. Moreover, the number of covariates p can be much larger than the number of observations n, especially for microarray and gene expression data in biological science (Nguyen and Rocke, 2002; Wu et al., 2009). When $p > n$, LDA may not perform very well. Bickel and Levina (2004) showed that LDA

can be asymptotically as bad as random guessing. The challenges with high dimensional data call for novel techniques to generalize LDA that can address the classification problems with $p > n$.

From the classification rule in (2.1), LDA involves estimating the covariance matrix $\mathbf{\Sigma}$. One major issue for LDA with $p > n$ is that the maximum likelihood estimate of $\mathbf{\Sigma}$ is singular hence the prediction performance suffers from high variance (Witten and Tibshirani, 2009; Shao et al., 2011). One way to resolve the singularity is to apply regularization when estimating the covariance matrix. A penalization method, called "Scout" (Witten and Tibshirani, 2009), is proposed to employ a shrinkage estimator for $\mathbf{\Sigma}^{-1}$. Specifically, the Scout method first computes the estimator of the precision matrix $\mathbf{\Sigma}^{-1}$. Then the LDA classification rule in (2.1) is constructed with $\mathbf{\Sigma}^{-1}$ plugged in by the shrinkage estimator $\hat{\mathbf{\Sigma}}^{-1}$. Rothman et al. (2008) proposed another approach based on the shrinkage estimator of $\mathbf{\Sigma}^{-1}$ called SPICE. Instead of penalizing on the inverse covariance matrix, LDA based on directly thresholding the sample covariance matrix $\boldsymbol{S}$ is proposed in Shao et al. (2011). There are also several methods considering the regularization on the mean $\boldsymbol{\mu}_k$ for improving the performance of LDA (Tibshirani et al., 2003; Guo et al., 2007).

Besides the idea of shrinkage estimator, another perspective is to reduce the number of variables such that one can directly perform the traditional LDA after variable screening. To select the effective variables among all the variables, a features annealed independence rule (FAIR) is proposed by Fan and Fan (2008). The FAIR method selects the important features according to their two sample t-test values. This method is easy to implement however ignores the correlation among variables. There are different variable screening method available in the literature. An overview of different variable selection techniques in high dimensional feature space can be found in Fan and Lv (2010).

Based on the LDA classification rule, Cai and Liu (2012) developed a direct estimation approach for sparse linear discriminant analysis using linear programming which is referred to as LPD. Rather than substituting the precision matrix by some regularized estimates as in Witten and Tibshirani (2009) and Rothman et al. (2008), they provide a shrinkage estimator

directly on $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$, where $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ is the difference between the means of the two classes. LPD performs well for data with large number of features and enjoys certain convergence properties. While the sparsity of the product $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ can be a relatively strong assumption. Even if $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$ are both sparse, their product may not be sparse. Moreover, the structure of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$ may also have some inherent patterns (Pavlenko et al., 2012).

In this chapter, I proposed using shrink estimates of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$ separately in the classification rule based on LDA. Since the proposed method employs the same classification rule as LDA, the proposed method enjoys the advantage of the linear boundary and ease of interpretation. With a large number of features, it is natural to assume that most features do not contribute to classification and are conditionally independent, that is, both $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$ can be sparse. By estimating $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$ separately, I will be able to maintain the sparse structure of both the inverse covariance matrix and the difference between the class means simultaneously. Furthermore, the estimation procedure of $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$ will be transformed to iteratively solving a graphical lasso (Friedman et al., 2007) and a lasso (Tibshirani, 1996) problem which can both be efficiently calculated.

The remainder of this chapter is organized as follows. Details of the proposed method is described in Section 2.2. To select the optimal tuning parameters in the proposed method, Section 2.3 adopts both misclassification error and extended Bayesian information criterion (EBIC) (Chen and Chen, 2008) as tuning parameters selection criteria. In Section 2.4, I conduct a simulation study to evaluate the proposed method. Two real data examples are used to elaborate the merits of the proposed method in Section 2.5. Section 2.6 concludes this chapter with some discussion.

2.2 Methodology

Consider a two-class classification problem with response $Y = 1$ or -1 corresponding to the two classes G_1 and G_2. Denote the predictor variables as $\boldsymbol{X} = \{X_1, \cdots, X_p\}$. The LDA assumes that the two classes share a same covariance structure denote as $\boldsymbol{\Sigma}$. I also denote the

means of $\boldsymbol{X}$ for the two classes G_1 and G_2 as $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively. Specifically,

$$G_1 : \boldsymbol{X}|Y = 1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), G_2 : \boldsymbol{X}|Y = -1 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \Re^p$. Suppose the observation data are $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{n_1}, \boldsymbol{x}_{n_1+1}, \cdots, \boldsymbol{x}_{n_1+n_2}$ with the first n_1 observations from G_1 and the last n_2 observations from G_2. Then the log-likelihood function of the data can be written as

$$\begin{aligned} L(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) =& n \ln |\boldsymbol{C}| - \sum_{k=1}^{2} \sum_{i \in G_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)' \boldsymbol{C} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \\ =& n \ln |\boldsymbol{C}| - \mathrm{tr}(\boldsymbol{C}\bar{\boldsymbol{S}}), \end{aligned} \tag{2.2}$$

up to some constant independent of parameters. Here $n = n_1 + n_2$, $\boldsymbol{C} = \boldsymbol{\Sigma}^{-1}$, $\bar{\boldsymbol{S}} = \sum_{k=1}^{2} \sum_{i \in G_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)'$. For a new data point $\boldsymbol{x}$, its classification label can be determined by the Bayes rule in (2.1), which depends on the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$. Equivalently, the classification rule will assign $\boldsymbol{x}$ to G_1 if

$$\ln \frac{\Pr(G_1|X = \boldsymbol{x})}{\Pr(G_2|X = \boldsymbol{x})} = \ln \frac{\pi_1}{\pi_2} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > 0, \tag{2.3}$$

where π_1 and π_2 are the prior probability of $\boldsymbol{x}$ to belong to G_1 and G_2, respectively. Otherwise $\boldsymbol{x}$ will be classified to G_2. The parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ can be estimated by maximizing the log-likelihood function in (2.2).

The log-ratio of conditional density functions in (2.3) involves the term $\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ with $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. When p is of the same size or larger than the sample size n, a regularized procedure is often needed to ensure proper estimation of $\boldsymbol{\Sigma}^{-1}$ or $\boldsymbol{\Sigma}$. As there are a large number of predictor variables, it is likely that some variables are irrelevant for classification (Fan and Fan, 2008). In this chapter, I propose a regularized approach in which separate regularizations are applied on $\boldsymbol{\Sigma}^{-1}$ and $\boldsymbol{\delta}$. Specifically, I consider the regularized log-likelihood function for

parameter estimation as follows:

$$\min_{\{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{C}\}} -n\ln|\boldsymbol{C}| + \sum_{k=1}^{2}\sum_{i\in G_k}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)'\boldsymbol{C}(\boldsymbol{x}_i-\boldsymbol{\mu}_k) + \lambda_1||\boldsymbol{C}||_1 + 2\lambda_2||\boldsymbol{\mu}_1-\boldsymbol{\mu}_2||_1, \tag{2.4}$$

where $||\cdot||_1$ is the l_1 norm, $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are two tuning parameters. Note that the constant term on λ_2 is for convenience purpose.

For the optimization in (2.4), the objective function involves two penalty terms $\boldsymbol{C}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. To efficiently estimate the parameters, I use an iterative procedure to solve the sub-optimization problem with respect to $\boldsymbol{C}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, respectively. Define $\boldsymbol{\delta}_h = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/2$, $\boldsymbol{\gamma} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Then we can write $\boldsymbol{\mu}_1 = \boldsymbol{\delta}_h + \boldsymbol{\gamma}$ and $\boldsymbol{\mu}_2 = \boldsymbol{\gamma} - \boldsymbol{\delta}_h$. The objective function in (2.4) can be rewritten as:

$$\begin{aligned}\min_{\{\boldsymbol{\delta}_h,\boldsymbol{\gamma},\boldsymbol{C}\}} & -n\ln|\boldsymbol{C}| + \sum_{i\in G_1}(\boldsymbol{x}_i-\boldsymbol{\delta}_h-\boldsymbol{\gamma})'\boldsymbol{C}(\boldsymbol{x}_i-\boldsymbol{\delta}_h-\boldsymbol{\gamma}) \\ & + \sum_{i\in G_2}(\boldsymbol{x}_i+\boldsymbol{\delta}_h-\boldsymbol{\gamma})'\boldsymbol{C}(\boldsymbol{x}_i+\boldsymbol{\delta}_h-\boldsymbol{\gamma}) + \lambda_1||\boldsymbol{C}||_1 + \lambda_2||\boldsymbol{\delta}_h||_1. \end{aligned} \tag{2.5}$$

Based on the objective function (2.5), it is easy to obtain the estimate of $\boldsymbol{\gamma}$ as:

$$\hat{\boldsymbol{\gamma}} = \bar{\boldsymbol{x}} + \frac{n_2 - n_1}{n}\boldsymbol{\delta}_h, \tag{2.6}$$

where $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$ is the overall mean. If $n_1 = n_2$, then the estimate of $\boldsymbol{\gamma}$ is just $\bar{\boldsymbol{x}}$. By plugging in $\hat{\boldsymbol{\gamma}}$ into the objective function (2.5), I get the minimization problem in terms of its profile likelihood function as:

$$\begin{aligned}\min_{\{\boldsymbol{\delta}_h,\boldsymbol{C}\}} & -n\ln|\boldsymbol{C}| + \sum_{i\in G_1}(\boldsymbol{x}_i-\frac{2n_2}{n}\boldsymbol{\delta}_h-\bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i-\frac{2n_2}{n}\boldsymbol{\delta}_h-\bar{\boldsymbol{x}}) \\ & + \sum_{i\in G_2}(\boldsymbol{x}_i+\frac{2n_1}{n}\boldsymbol{\delta}_h-\bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i+\frac{2n_1}{n}\boldsymbol{\delta}_h-\bar{\boldsymbol{x}}) + \lambda_1||\boldsymbol{C}||_1 + \lambda_2||\boldsymbol{\delta}_h||_1. \end{aligned} \tag{2.7}$$

Note that the new objective function (2.7) has the form of profile likelihood since $\boldsymbol{\gamma}$ is plugged in by its maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$. By applying the penalization on both $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$,

the proposed method is able to estimate the sparse structures in $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$ simultaneously.

For efficiently estimating unknown parameters $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$ in (2.7), we can iteratively solve for $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$ until they both converge. For a given $\boldsymbol{\delta}_h$, the minimization problem (2.7) with respect to $\boldsymbol{C}$ is equivalent to:

$$\min_{\boldsymbol{C}} -\ln|\boldsymbol{C}| + \mathrm{tr}(\boldsymbol{C}\tilde{\boldsymbol{S}}) + \lambda_1||\boldsymbol{C}||_1, \tag{2.8}$$

where $\tilde{\boldsymbol{S}} = \frac{1}{n}(\sum_{i\in G_1}(\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}}) + \sum_{i\in G_2}(\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}}))$. Clearly, it has the same formulation as the objective function in the graphical lasso (Friedman et al., 2007). When $\boldsymbol{C}$ is given, the minimization problem (2.7) with respect to $\boldsymbol{\delta}_h$ is equivalent to:

$$\begin{aligned}\min_{\boldsymbol{\delta}_h} &\sum_{i\in G_1}(\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}}) \\ &+ \sum_{i\in G_2}(\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}}) + \lambda_2||\boldsymbol{\delta}_h||_1 \\ \propto &(\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\delta}_h)'(\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\delta}_h) + \lambda_2||\boldsymbol{\delta}_h||_1,\end{aligned} \tag{2.9}$$

where $\tilde{\boldsymbol{y}} = \frac{n}{2n_1n_2}\boldsymbol{C}^{1/2}(\sum_{i\in G_1}\boldsymbol{x}_i - n_1\bar{\boldsymbol{x}})$, $\tilde{X} = \boldsymbol{C}^{1/2}$. A detailed derivation of (2.9) can be found in Appendix A. The objective function (2.9) becomes a lasso regression problem (Tibshirani, 1996). Thus solving the optimization in (2.7) can be decomposed by iteratively solving a graphical lasso (Friedman et al., 2007) problem for $\boldsymbol{C}$ and solving a lasso (Tibshirani, 1996) problem for $\boldsymbol{\delta}_h$. The algorithm to estimate $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$ is described as follows:

Algorithm 1

Step 0: *Set the initial value of* $\boldsymbol{\delta}_h$ *as* $\hat{\boldsymbol{\delta}}_h = \boldsymbol{\delta}_0$.

Step 1: *Given* $\boldsymbol{\delta}_h = \hat{\boldsymbol{\delta}}_h$, *solve for* $\hat{\boldsymbol{C}}$ *by minimizing (2.8) using the graphical lasso approach.*

Step 2: *Given* $\boldsymbol{C} = \hat{\boldsymbol{C}}$, *solve for* $\hat{\boldsymbol{\delta}}_h$ *by minimizing (2.9) using the lasso approach.*

Step 3: *Go back to* Step 1 *and repeat* Step 1 *and* Step 2 *until both* $\hat{\boldsymbol{C}}$ *and* $\hat{\boldsymbol{\delta}}_h$ *converge.*

The convergence criteria are $||\hat{\boldsymbol{C}}_t - \hat{\boldsymbol{C}}_{t-1}||_F^2 < \tau_1$ and $||\hat{\boldsymbol{\delta}}_{h,t} - \hat{\boldsymbol{\delta}}_{h,t-1}||_2^2 < \tau_2$ where $\hat{\boldsymbol{C}}_t$, $\hat{\boldsymbol{\delta}}_{h,t}$,

$\hat{\boldsymbol{C}}_{t-1}$ and $\hat{\boldsymbol{\delta}}_{h,t-1}$ are the estimates of $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$ in the tth and $(t-1)$th iteration, τ_1 and τ_2 are two pre-selected small quantities. Note that both objective functions (2.8) and (2.9) are in convex forms such that convergence of the algorithm is guaranteed (Cancès et al., 2011). For the choice of initial value $\boldsymbol{\delta}_0$, I set at $(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)/2$ where $\bar{\boldsymbol{x}}_k$, $k = 1, 2$, is the sample mean for the kth class. After $\hat{\boldsymbol{\delta}}_h$ is estimated through Algorithm 1, $\hat{\boldsymbol{\gamma}}$ can be calculated as (2.6) plugging in $\hat{\boldsymbol{\delta}}_h$ for $\boldsymbol{\delta}_h$. Then $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ can be estimated by $\hat{\boldsymbol{\mu}}_1 = \hat{\boldsymbol{\delta}}_h + \hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\delta}}_h$ respectively. Two-class classification can be performed by plugging in $\hat{\boldsymbol{C}}$, $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ into the classification rule (2.3).

2.3 Tuning Parameters Selection

There are two tuning parameters λ_1 and λ_2 in (2.7), where λ_1 is the shrinkage parameter on $\boldsymbol{C}$ and λ_2 is the shrinkage parameter on $\boldsymbol{\delta}_h$. For choosing the optimal tuning parameters, I consider two criteria: the misclassification error (ME) and the extended Bayesian information criterion (EBIC) (Chen and Chen, 2008; Foygel and Drton, 2010). Here I adopt the commonly used cross validation (Davison and Hall, 1992) for calculating the misclassification error. Under the cross validation context, with a K fold $J_1, \cdots, J_K$, the misclassification error can be calculated as

$$Error(\lambda_1, \lambda_2) = \sum_{j}^{K} \sum_{i \in J_k} I(\hat{y}_i \neq G(\boldsymbol{x}_i)), \tag{2.10}$$

where $I(\cdot)$ is an indicator function, $\hat{y}_i$ is the estimated class label for $\boldsymbol{x}_i$, and $G(\boldsymbol{x}_i)$ is the true class label for $\boldsymbol{x}_i$. When using EBIC as the tuning parameter selection criterion. The EBIC can be computed as follows:

$$EBIC(\lambda_1, \lambda_2) = -n \ln |\hat{\boldsymbol{C}}| + tr(\hat{\boldsymbol{C}}\tilde{\boldsymbol{S}}) + v(\hat{\boldsymbol{\delta}}_h) \ln n + 2\gamma_1 \ln \tau(v(\hat{\boldsymbol{\delta}}_h)) + |E| \ln n + 4|E|\gamma_2 \ln p,$$

where $\hat{\boldsymbol{\delta}}_h$ and $\hat{\boldsymbol{C}}$ are the estimates of $\boldsymbol{\delta}_h$ and $\boldsymbol{C}$, $v(\hat{\boldsymbol{\delta}}_h)$ and $|E|$ are the degrees of freedom of $\hat{\boldsymbol{\delta}}_h$ and $\hat{\boldsymbol{C}}$, respectively, $\tau(v(\hat{\boldsymbol{\delta}}_h)) = \binom{p}{v(\hat{\boldsymbol{\delta}}_h)}$, $\gamma_1, \gamma_2 \in [0, 1]$ are two penalizing parameters. Different from the misclassification error computed by (2.10), the calculation of EBIC does

not need the use of cross validation.

I like to remark that the use of the misclassification error criterion allows the proposed method to concentrate more on classification accuracy. When p becomes large, the calculation of misclassification error using cross validation can be computationally expensive. Alternatively, the EBIC criterion can be much faster to compute. In addition to its fast calculation, the EBIC also enjoys consistency properties under certain conditions of p and n such as $p = O(n^{\kappa})$ (Chen and Chen, 2008) where κ is a constant. Based on a pre-chosen set of combinations of λ_1, λ_2, I select the optimal tuning parameters as the one minimizing the misclassification error or the EBIC. The performances of both selection criteria will be illustrated in Sections 2.4 and 2.5.

2.4 Simulation

In this section, I evaluate the performance of the proposed method under the consideration of different inverse covariance matrix $\boldsymbol{C}$ and mean difference $\boldsymbol{\delta}_h$ with various levels of sparsity.

For the inverse covariance matrix $\boldsymbol{C} = (c_{ij})_{p\times p}$, I consider five different covariance structures:

- Model 1. Independent (I) setting: $c_{ij} = 1$ if $i = j$ and 0 otherwise;
- Model 2. AR(1) setting, $c_{ij} = 1$ if $i = j$, $c_{ij} = 0.45$ if $|i - j| = 1$ and 0 otherwise;
- Model 3. AR(2) setting, $c_{ij} = 1$ if $i = j$, $c_{ij} = 0.45$ if $|i - j| = 1$, $c_{ij} = 0.25$ if $|i - j| = 2$ and 0 otherwise;
- Model 4. Permuted AR(1) (PAR(1)) setting: generate $\boldsymbol{C}_0$ under the AR(1) setting and a permutation matrix $\boldsymbol{P}$, then set $\boldsymbol{C} = \boldsymbol{P}'\boldsymbol{C}_0\boldsymbol{P}$;
- Model 5. Permuted AR(2) (PAR(2)) setting: generate $\boldsymbol{C}_0$ under the AR(2) setting and a permutation matrix $\boldsymbol{P}$, then set $\boldsymbol{C} = \boldsymbol{P}'\boldsymbol{C}_0\boldsymbol{P}$.

For the mean difference $\boldsymbol{\delta}_h$, I consider three different levels of sparsity as follows. The $\boldsymbol{\mu}_1$ is set as a vector with all elements being zeros. Then three cases of $\boldsymbol{\mu}_2$ are considered. (S1): 25%

of the elements in $\boldsymbol{\mu}_2$ are zeros; (S2): 50% of the elements in $\boldsymbol{\mu}_2$ are zeros; (S3): 75% of the elements in $\boldsymbol{\mu}_2$ are zeros. The positions of the zero elements in $\boldsymbol{\mu}_2$ are randomly distributed throughout all positions. For non-zero elements in $\boldsymbol{\mu}_2$, their values are independently generated from the uniform distribution $U[0, 1]$. Note that even when both $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$ are sparse, their product $\boldsymbol{C}\boldsymbol{\delta}_h$ may not be sparse. For example, in the case of S2 that 50% of the elements in $\boldsymbol{\delta}_h$ be set to 0, there can be 50% of the elements in $\boldsymbol{C}\boldsymbol{\delta}_h$ having non-zero values under Model 1. While there can be more than 90% of the elements in $C\boldsymbol{\delta}_h$ having non-zero values under Model 2 and Model 4, and all elements in $C\boldsymbol{\delta}_h$ will be non-zero under Model 3 and Model 5. It shows that the sparsity of both $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$ may not imply $\boldsymbol{C}\boldsymbol{\delta}_h$ being sparse.

For each setting of $\boldsymbol{C}$ and $\boldsymbol{\delta}_h$, the number of predictor variables is set at $p = 25, 50, 100$ respectively. To obtain a training data set, I generate $n_1 = 50$ observations for G_1 from $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $n_2 = 50$ observations for G_2 from $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where $\boldsymbol{C} \triangleq \boldsymbol{\Sigma}^{-1}$. The same generation procedure is used to generate the test data set, which is used for evaluating the classification accuracy for different methods in comparison.

Let us denote the proposed method under ME and EBIC criteria as SLDA-ME and SLDA-EBIC respectively. I compare the proposed method with several existing methods including the generalized LDA (GLDA) with $\boldsymbol{C}$ estimated by the generalized inverse covariance matrix (Howland and Park, 2004), the Scout method (Witten and Tibshirani, 2009), the features annealed independence rule (FAIR) (Fan and Fan, 2008) and the linear programming discriminant (LPD) rule (Cai and Liu, 2012). Here the GLDA is used as a benchmark classification method for comparison. The Scout method (Witten and Tibshirani, 2009) applies the l_1 penalty on $\boldsymbol{C}$ and estimates the precision matrix as:

$$\hat{\boldsymbol{C}} = \arg\max_{\boldsymbol{C}} \{\log\det \boldsymbol{C} - \mathrm{tr}(\boldsymbol{S}_n \boldsymbol{C}) - \lambda ||\boldsymbol{C}||\},$$

where $\boldsymbol{S}_n$ is the pooled sample covariance matrix. Then the estimate $\hat{\boldsymbol{C}}$ is plugged into the

LDA rule (2.1) for classification. The LPD method is to estimate $\tilde{\boldsymbol{\beta}} = \boldsymbol{C}\boldsymbol{\delta}$ directly by

$$\hat{\tilde{\boldsymbol{\beta}}} = \arg\min_{\beta^* \in \Re^p} \{|\tilde{\boldsymbol{\beta}}|_1 \text{ subject to } |\boldsymbol{S}_n\tilde{\boldsymbol{\beta}} - (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)|_\infty \leq \lambda_n\}. \tag{2.11}$$

From (2.11) it is clear that LPD is not a likelihood based estimator. The FAIR method (Fan and Fan, 2008) is to reduce the effective number of variables used in LDA for classification. By performing individual two sample t-test for each variable, the FAIR selects m_0 variables that have the largest t statistics, where m_0 is determined as

$$\hat{m}_0 = \arg\max_{1 \leq m \leq p} \frac{[\sum_{j=1}^{m} \hat{\alpha}_j^2 + m(n_1 - n_2)/(n_1 n_2)]^2}{nm/(n_1 n_2) + \sum_{j=1}^{m} \hat{\alpha}_j^2},$$

where $\hat{\alpha}_j = \hat{\mu}_{1j} - \hat{\mu}_{2j}$ with $\hat{\mu}_{1j}$ and $\hat{\mu}_{2j}$ being the samples means of the jth variable for G_1 and G_2. The traditional LDA is performed afterwards based on the selected m_0 variables.

I report the comparison results under five simulation models of $\boldsymbol{C}$ with all the three degrees of sparsity for $\boldsymbol{\delta}_h$ under different p settings. The data generation procedure for training and test data is repeated 50 times for each simulation setting. The 10-fold cross validation is used for tuning parameters selection under SLDA-ME, the Scout procedure and the LPD. For the SLDA-EBIC, $\gamma_1 = \gamma_2 = 0.5$ are used. The actual number of misclassified cases out of the test data is used to compare the performances of the proposed methods with other approaches. Table 2.1 reports the averaged misclassification rates as well as their standard deviations in parenthesis for the six methods in comparison. Detailed boxplots regarding the misclassification errors are displayed in Appendix B from Figure 1 to Figure 15.

From the results in Table 2.1, SLDA-ME and SLDA-EBIC generally perform better than LPD, FAIR and Scout in terms of misclassification error. In the case of Model 1 with $\boldsymbol{C} = \boldsymbol{I}$, the misclassification errors of the proposed methods are comparable to those of Scout, FAIR, and LPD. But for the other models of $\boldsymbol{C}$, it is clear that the proposed methods give smaller misclassification errors than other methods. The advantage of the proposed methods over the

other methods become more significant as the dimension p increases. Moreover, as the degree of sparsity of $\boldsymbol{\delta_h}$ gets larger from S1 to S3, the SLDA-ME appears to outperform SLDA-EBIC gradually. Specifically, under a less sparse $\boldsymbol{\delta_h}$ case S1 that 25% of the elements in $\boldsymbol{\delta_h}$ are set to zero, the performance of SLDA-EBIC is comparable to SLDA-ME. While under the cases of S2 and S3 that 50% or 75% of the elements in $\boldsymbol{\delta_h}$ are zeros, the SLDA-ME performs better than SLDA-EBIC. Note that SLDA-ME uses the misclassification error under cross validation for choosing the tuning parameter. This can be a possible explanation that SLDA-ME can yield a smaller ME in comparison with SLDA-EBIC. But it is worth pointing out that the SLDA-ME requires more computation because of cross validation while SLDA-EBIC does not.

Recall that under the case S2, there are more than 90% of the elements in $\boldsymbol{C\delta}_h$ having non-zero values under Model 2 and Model 4, and all elements in $\boldsymbol{C\delta}_h$ being non-zero under Model 3 and Model 5. For these scenarios, one can see that the proposed methods perform better than LPD which requires the assumption for the sparsity of $\boldsymbol{C\delta}_h$. A possible explanation is that when $\boldsymbol{C\delta}_h$ is not sparse, the l_1 penalization on $\boldsymbol{C\delta}_h$ in LPD could force many elements in $\boldsymbol{C\delta}_h$ to be estimated as 0, which can cause a larger ME compared with the proposed methods. Even when $\boldsymbol{C\delta}_h$ is sparse in the case of S1 with Model 1 for $\boldsymbol{C}$, the proposed methods can outperform LPD especially for a large $p = 100$.

To obtain more insight of the proposed methods, I further examine the estimation accuracy of $\boldsymbol{\beta} = \boldsymbol{C\delta}_h$ since the classification rule depends on $\boldsymbol{\beta}$. By denoting the estimate of $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, I evaluate the accuracy of $\hat{\boldsymbol{\beta}}$ by

$$D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2.$$

Let us denote by $\hat{\boldsymbol{\beta}}_{ME}$, $\hat{\boldsymbol{\beta}}_{EBIC}$ and $\hat{\boldsymbol{\beta}}_{LPD}$ the estimates of $\boldsymbol{\beta}$ from SLDA-ME, SLDA-EBIC and LPD, respectively. Table 2.2 reports the results of the estimation accuracy of $\boldsymbol{\beta}$ by its distance measure $D(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$. Table 2.2 also reports the original $||\boldsymbol{\beta}||_2$ values. From the results in Table 2.2, we can see that under most cases $D(\hat{\boldsymbol{\beta}}_{ME}, \boldsymbol{\beta})$ and $D(\hat{\boldsymbol{\beta}}_{EBIC}, \boldsymbol{\beta})$ are smaller than $D(\hat{\boldsymbol{\beta}}_{LPD}, \boldsymbol{\beta})$. It means that the $\hat{\boldsymbol{\beta}}_{LPD}$ has the larger distance to the true $\boldsymbol{\beta}$ in comparison with the proposed methods, despite the degrees of sparsity of $\boldsymbol{\delta}_h$. Such an observation agrees

Table 2.1: Averaged misclassification rates and standard errors (in parenthesis) from 100 replications

Data Generation Setting			SLDA-ME	SLDA-EBIC	GLDA	Scout	FAIR	LPD
S1	I	p=25	11.68 (0.37)	10.60 (0.31)	14.20 (0.42)	12.54 (0.35)	9.60 (0.30)	13.16 (0.33)
		p=50	6.56 (0.30)	5.48 (0.24)	13.46 (0.39)	9.14 (0.33)	5.14 (0.23)	10.24 (0.38)
		p=100	1.96 (0.14)	1.06 (0.11)	32.66 (0.77)	5.10 (0.26)	1.06 (0.11)	5.56 (0.33)
	AR(1)	p=25	11.90 (0.38)	10.86 (0.35)	13.40 (0.35)	29.22 (0.61)	34.94 (0.65)	12.48 (0.29)
		p=50	3.32 (0.19)	2.78 (0.14)	7.32 (0.33)	14.98 (0.69)	21.48 (0.65)	6.38 (0.25)
		p=100	1.14 (0.12)	0.30 (0.06)	31.42 (0.91)	8.04 (0.45)	11.58 (0.37)	6.02 (0.33)
	AR(2)	p=25	7.46 (0.22)	9.16 (0.28)	8.04 (0.25)	16.44 (0.51)	18.12 (0.47)	10.80 (0.38)
		p=50	0.80 (0.10)	1.18 (0.10)	2.54 (0.17)	4.38 (0.31)	4.30 (0.22)	3.62 (0.21)
		p=100	1.34 (0.12)	0.96 (0.09)	30.00 (0.86)	7.50 (0.31)	2.94 (0.20)	7.90 (0.35)
	PAR(1)	p=25	14.82 (0.34)	14.58 (0.36)	16.00 (0.38)	31.04 (0.58)	39.9 (0.49)	17.68 (0.50)
		p=50	2.60 (0.18)	2.48 (0.15)	5.80 (0.28)	12.72 (0.50)	19.28 (0.71)	4.74 (0.27)
		p=100	1.69 (0.13)	0.99 (0.08)	30.63 (0.70)	14.81 (0.60)	15.75 (0.40)	8.35 (0.33)
	PAR(2)	p=25	6.88 (0.27)	9.24 (0.34)	7.50 (0.32)	14.34 (0.53)	16.82 (0.42)	8.24 (0.35)
		p=50	5.56 (0.23)	6.84 (0.26)	8.84 (0.34)	14.76 (0.43)	15.18 (0.40)	7.78 (0.30)
		p=100	0.68 (0.08)	0.30 (0.05)	28.14 (0.83)	5.44 (0.38)	2.04 (0.16)	7.10 (0.41)
S2	I	p=25	16.82 (0.41)	15.92 (0.43)	23.84 (0.45)	13.36 (0.34)	20.80 (0.34)	9.48 (0.34)
		p=50	16.38 (0.36)	15.40 (0.51)	29.46 (0.50)	12.22 (0.37)	19.88 (0.31)	10.14 (0.32)
		p=100	4.56 (0.19)	3.96 (0.19)	29.52 (0.89)	3.06 (0.34)	12.84 (0.36)	3.95 (0.23)
	AR(1)	p=25	15.86 (0.41)	16.12 (0.38)	25.50 (0.44)	38.00 (0.51)	30.02 (0.50)	13.28 (0.43)
		p=50	10.76 (0.36)	11.78 (0.28)	24.68 (0.56)	35.74 (0.39)	26.82 (0.44)	10.96 (0.39)
		p=100	2.00 (0.11)	2.16 (0.18)	28.40 (0.47)	16.30 (0.43)	15.20 (0.28)	5.62 (0.26)
	AR(2)	p=25	9.64 (0.27)	13.10 (0.41)	20.18 (0.34)	22.78 (0.46)	22.64 (0.42)	8.54 (0.33)
		p=50	10.20 (0.31)	12.34 (0.34)	24.46 (0.40)	20.02 (0.38)	20.90 (0.41)	10.92 (0.43)
		p=100	2.12 (0.16)	3.02 (0.21)	29.66 (0.54)	6.60 (0.28)	14.76 (0.28)	7.46 (0.29)
	PAR(1)	p=25	18.02 (0.42)	19.26 (0.51)	27.66 (0.44)	38.74 (0.41)	29.42 (0.45)	15.6 (0.41)
		p=50	6.34 (0.24)	7.22 (0.24)	21.70 (0.34)	30.08 (0.36)	24.10 (0.55)	7.88 (0.28)
		p=100	3.42 (0.15)	4.44 (0.27)	30.42 (0.63)	23.18 (0.40)	18.26 (0.33)	8.48 (0.37)
	PAR(2)	p=25	10.74 (0.33)	14.34 (0.41)	21.94 (0.39)	24.54 (0.38)	23.84 (0.39)	10.66 (0.41)
		p=50	9.52 (0.32)	12.54 (0.35)	23.56 (0.37)	21.82 (0.42)	21.52 (0.43)	11.02 (0.35)
		p=100	2.12 (0.17)	3.48 (0.21)	27.80 (0.47)	5.58 (0.21)	15.00 (0.26)	7.64 (0.31)
S3	I	p=25	27.96 (0.48)	26.42 (0.50)	29.48 (0.52)	27.22 (0.47)	25.14 (0.51)	27.50 (0.60)
		p=50	22.56 (0.39)	21.50 (0.43)	28.62 (0.52)	23.40 (0.53)	21.30 (0.38)	23.12 (0.50)
		p=100	10.54 (0.36)	7.96 (0.26)	39.84 (0.78)	10.42 (0.41)	6.80 (0.26)	11.82 (0.40)
	AR(1)	p=25	33.88 (0.59)	34.80 (0.47)	34.30 (0.57)	40.52 (0.42)	45.28 (0.46)	35.06 (0.57)
		p=50	22.26 (0.51)	22.60 (0.49)	28.82 (0.48)	32.34 (0.63)	41.92 (0.54)	26.38 (0.53)
		p=100	11.98 (0.37)	12.94 (0.35)	40.72 (0.83)	23.08 (0.51)	36.92 (0.58)	21.38 (0.65)
	AR(2)	p=25	20.34 (0.45)	23.36 (0.44)	22.24 (0.47)	26.02 (0.43)	31.62 (0.43)	25.68 (0.59)
		p=50	21.56 (0.46)	25.28 (0.55)	27.18 (0.48)	25.90 (0.46)	32.46 (0.40)	28.08 (0.50)
		p=100	15.68 (0.36)	16.50 (0.37)	39.04 (0.59)	20.16 (0.48)	22.26 (0.40)	24.02 (0.75)
	PAR(1)	p=25	21.96 (0.40)	22.28 (0.37)	23.82 (0.46)	32.54 (0.65)	43.46 (0.51)	24.38 (0.56)
		p=50	18.18 (0.42)	18.60 (0.33)	25.42 (0.58)	27.20 (0.50)	40.02 (0.56)	22.78 (0.47)
		p=100	16.31 (0.46)	16.50 (0.38)	44.19 (0.70)	26.26 (0.57)	38.90 (0.48)	24.58 (0.60)
	PAR(2)	p=25	27.28 (0.49)	28.68 (0.51)	26.28 (0.51)	31.46 (0.50)	36.20 (0.58)	29.30 (0.50)
		p=50	24.04 (0.46)	25.56 (0.48)	27.88 (0.53)	28.86 (0.45)	32.40 (0.46)	27.52 (0.58)
		p=100	17.38 (0.40)	19.72 (0.37)	41.44 (0.61)	21.78 (0.56)	27.72 (0.51)	25.30 (0.66)

with the result shown in Table 2.1, where the SLDA-ME and SLDA-EBIC perform better than LPD. As can be seen in (2.3), the accuracy of the estimated $\hat{\boldsymbol{\beta}}$ can directly affect the performance of the classification error.

Moreover, I also examine the estimation accuracy of $\boldsymbol{C}$. Here I use the estimation error percentage of the inverse covariance matrix as

$$\frac{||\hat{\boldsymbol{C}} - \boldsymbol{C}||_F^2}{||\boldsymbol{C}||_F^2},$$

where $||\cdot||_F$ is the Frobenius norm (Golub and Van Loan, 1996). Table 2.3 reports the average estimation error percentage along with the standard deviations for the inverse covariance matrix $\boldsymbol{C}$. From Table 2.3, we can see that the SLDA-EBIC provides more stable estimation of $\boldsymbol{C}$ compared with the SLDA-ME. Under each model, the estimation accuracy of $\boldsymbol{C}$ from SLDA-EBIC remains around a certain percentage level across different degrees of sparsity for $\boldsymbol{\delta}_h$ and different p levels. While for SLDA-ME, the estimation error percentage varies across different degrees of sparsity for $\boldsymbol{\delta}_h$ and differs significantly among different p levels. Regarding the estimation accuracy of $\boldsymbol{C}$, the SLDA-ME outperforms SLDA-EBIC when p is relatively small, while SLDA-EBIC gives better accuracy when $p = 100$. Moreover, the results of Table 2.3 suggest that SLDA-EBIC yields a more accurate estimate of $\boldsymbol{C}$ when the degree of sparsity for $\boldsymbol{C}$ is higher. Note that Table 2.1 suggests that SLDA-ME can give more accurate classification accuracy than SLDA-EBIC under most settings. Since the classification accuracy does not solely depend on the estimation of $\boldsymbol{C}$ but the product $\boldsymbol{\beta} = \boldsymbol{C}\boldsymbol{\delta}_h$. Even though SLDA-EBIC can yield a more accurate estimate of $\boldsymbol{C}$ under large p, SLDA-ME may still outperform SLDA-EBIC in classification.

2.5 Real Data Examples

To further illustrate the classification performance of the proposed methods, I apply the proposed methods for classification on two real data examples: the Leukemia data (Golub, et al., 1999) and the Prostate Cancer data (Singh et al., 2002). The proposed methods SLDA-ME

Table 2.2: Average and standard errors in parenthesis of the estimation accuracy for $\boldsymbol{\beta}$ from 100 replications. Last column shows the true $||\boldsymbol{\beta}||_2$ values

Data Generation Setting			$\|\|\hat{\boldsymbol{\beta}}_{ME}-\boldsymbol{\beta}\|\|_2$	$\|\|\hat{\boldsymbol{\beta}}_{EBIC}-\boldsymbol{\beta}\|\|_2$	$\|\|\hat{\boldsymbol{\beta}}_{LPD}-\boldsymbol{\beta}\|\|_2$	$\|\|\boldsymbol{\beta}\|\|_2$
S1	I	p=25	2.65 (0.06)	2.83 (0.03)	3.41 (0.20)	7.46
		p=50	4.04 (0.07)	4.59 (0.04)	6.82 (0.41)	11.62
		p=100	12.15 (0.77)	10.16 (0.06)	21.61 (0.79)	26.00
	AR(1)	p=25	5.68 (0.13)	7.69 (0.05)	6.93 (0.44)	14.40
		p=50	12.06 (0.31)	17.71 (0.08)	23.12 (1.51)	33.71
		p=100	22.38 (0.58)	34.08 (0.10)	46.46 (1.88)	65.27
	AR(2)	p=25	9.43 (0.23)	16.32 (0.05)	6.42 (0.37)	24.36
		p=50	21.56 (0.70)	45.56 (0.09)	29.52 (1.79)	68.36
		p=100	30.63 (0.76)	56.62 (0.12)	60.54 (2.45)	85.18
	PAR(1)	p=25	4.17 (0.12)	5.60 (0.04)	5.62 (0.44)	10.36
		p=50	13.10 (0.31)	18.55 (0.08)	27.61 (1.87)	35.51
		p=100	29.65 (0.63)	33.69 (0.13)	46.37 (1.30)	67.52
	PAR(2)	p=25	9.02 (0.29)	15.92 (0.05)	9.04 (0.52)	24.39
		p=50	14.11 (0.35)	23.62 (0.06)	27.51 (1.60)	35.21
		p=100	31.22 (0.67)	65.66 (0.11)	76.50 (2.74)	98.66

Data Generation Setting			$\|\|\hat{\boldsymbol{\beta}}_{ME}-\boldsymbol{\beta}\|\|_2$	$\|\|\hat{\boldsymbol{\beta}}_{EBIC}-\boldsymbol{\beta}\|\|_2$	$\|\|\hat{\boldsymbol{\beta}}_{LPD}-\boldsymbol{\beta}\|\|_2$	$\|\|\boldsymbol{\beta}\|\|_2$
S2	I	p=25	2.54 (0.05)	2.50 (0.03)	2.02 (0.05)	6.51
		p=50	3.01 (0.07)	2.99 (0.05)	4.89 (0.18)	6.82
		p=100	5.93 (0.10)	5.06 (0.09)	8.20 (0.24)	13.65
	AR(1)	p=25	2.63 (0.06)	3.30 (0.03)	2.21 (0.08)	5.30
		p=50	8.67 (0.20)	11.86 (0.04)	11.68 (0.57)	18.83
		p=100	15.64 (0.33)	15.99 (0.08)	25.02 (0.65)	34.00
	AR(2)	p=25	8.01 (0.13)	9.29 (0.04)	6.08 (0.47)	15.83
		p=50	16.16 (0.26)	19.50 (0.10)	21.27 (1.35)	31.54
		p=100	23.70 (0.50)	28.01 (0.09)	32.93 (0.57)	41.33
	PAR(1)	p=25	2.78 (0.06)	3.53 (0.03)	3.27 (0.21)	5.70
		p=50	9.04 (0.16)	11.24 (0.04)	16.77 (1.27)	19.40
		p=100	13.56 (1.05)	15.77 (0.04)	21.50 (0.81)	21.69
	PAR(2)	p=25	4.87 (0.10)	5.50 (0.05)	4.40 (0.19)	9.17
		p=50	15.03 (0.29)	17.68 (0.08)	19.44 (1.06)	28.98
		p=100	24.69 (0.45)	37.20 (0.12)	38.26 (0.80)	51.96

Data Generation Setting			$\|\|\hat{\boldsymbol{\beta}}_{ME}-\boldsymbol{\beta}\|\|_2$	$\|\|\hat{\boldsymbol{\beta}}_{EBIC}-\boldsymbol{\beta}\|\|_2$	$\|\|\hat{\boldsymbol{\beta}}_{LPD}-\boldsymbol{\beta}\|\|_2$	$\|\|\boldsymbol{\beta}\|\|_2$
S3	I	p=25	0.99 (0.02)	0.92 (0.01)	1.37 (0.16)	2.16
		p=50	1.84 (0.03)	1.71 (0.02)	2.27 (0.26)	3.78
		p=100	5.68 (0.38)	4.57 (0.04)	6.15 (0.22)	10.90
	AR(1)	p=25	1.17 (0.03)	1.27 (0.02)	2.07 (0.12)	2.28
		p=50	3.28 (0.06)	3.45 (0.03)	4.75 (0.33)	6.29
		p=100	7.47 (0.14)	7.63 (0.04)	13.17 (0.73)	14.41
	AR(2)	p=25	2.94 (0.08)	4.12 (0.02)	2.73 (0.20)	6.53
		p=50	4.30 (0.10)	5.32 (0.03)	6.02 (0.32)	8.07
		p=100	10.55 (0.26)	12.57 (0.04)	13.85 (0.72)	19.03
	PAR(1)	p=25	2.25 (0.06)	2.83 (0.03)	3.48 (0.26)	5.45
		p=50	3.87 (0.06)	4.01 (0.03)	7.28 (0.71)	7.56
		p=100	5.02 (0.27)	5.05 (0.23)	8.91 (0.66)	9.35
	PAR(2)	p=25	2.13 (0.06)	2.72 (0.02)	2.87 (0.15)	4.20
		p=50	3.85 (0.09)	4.60 (0.03)	7.56 (0.53)	7.19
		p=100	8.35 (0.15)	9.83 (0.04)	11.75 (0.32)	14.92

Table 2.3: Average and standard errors in parenthesis of the estimation accuracy for the inverse covariance matrix $\boldsymbol{C}$ from 100 replications

		S1		S2		S3	
		SLDA-ME	SLDA-EBIC	SLDA-ME	SLDA-EBIC	SLDA-ME	SLDA-EBIC
I	p=25	11.29% (0.90%)	4.12% (0.06%)	6.05% (0.45%)	4.89% (0.12%)	11.72% (0.84%)	3.91% (0.06%)
	p=50	13.79% (1.25%)	4.03% (0.06%)	8.17% (0.94%)	5.02% (0.14%)	15.28% (0.86%)	4.09% (0.04%)
	p=100	17.99% (1.32%)	4.25% (0.03%)	17.08% (0.79%)	4.95% (0.11%)	22.29% (0.96%)	4.24% (0.03%)
AR(1)	p=25	9.07% (0.41%)	16.55% (0.10%)	9.74% (0.43%)	15.28% (0.13%)	9.59% (0.48%)	16.22% (0.10%)
	p=50	18.24% (1.70%)	16.34% (0.07%)	8.72% (0.31%)	16.03% (0.14%)	13.73% (1.26%)	16.45% (0.06%)
	p=100	21.02% (1.70%)	16.04% (0.11%)	20.28% (1.36%)	16.44% (0.09%)	18.10% (1.20%)	16.21% (0.05%)
AR(2)	p=25	12.79% (0.27%)	30.35% (0.09%)	14.75% (0.45%)	28.64% (0.17%)	12.77% (0.37%)	30.46% (0.08%)
	p=50	31.89% (1.58%)	30.83% (0.06%)	26.37% (1.79%)	27.73% (0.18%)	19.99% (1.17%)	30.87% (0.05%)
	p=100	38.08% (2.01%)	30.96% (0.09%)	32.09% (2.30%)	30.03% (0.13%)	23.93% (1.00%)	30.93% (0.03%)
PAR(1)	p=25	10.85% (0.54%)	16.31% (0.10%)	11.55% (0.54%)	16.02% (0.12%)	8.66% (0.33%)	16.04% (0.10%)
	p=50	14.99% (1.37%)	16.44% (0.08%)	12.53% (0.81%)	16.13% (0.13%)	12.72% (0.14%)	16.07% (0.08%)
	p=100	21.38% (1.43%)	15.91% (0.04%)	22.47% (1.09%)	18.23% (0.20%)	20.29% (1.86%)	16.05% (0.05%)
PAR(2)	p=25	13.46% (0.33%)	30.31% (0.07%)	14.09% (0.46%)	30.32% (0.12%)	13.38% (0.32%)	30.50% (0.08%)
	p=50	27.12% (1.66%)	30.85% (0.06%)	22.20% (1.28%)	30.24% (0.08%)	22.96% (1.26%)	30.87% (0.06%)
	p=100	47.16% (2.74%)	31.07% (0.04%)	31.29% (2.20%)	31.97% (0.08%)	26.05% (1.69%)	31.05% (0.04%)

and SLDA-EBIC are compared with GLDA, Scout, FAIR, and LPD.

2.5.1 Leukemia Data

The Leukemia data is first studied in Golub, et al. (1999) and is available at http://www.broad institute.org/cgi-bin/cancer/datasets.cgi. This data set contains samples from two classes of leukemia patients, acute lymphoblastic leukemia (ALL) patients and acute myelogenous leukemia (AML) patients. This data contains 72 gene samples with the same 7129 related genes collected from 72 patients either with ALL or AML. Among all the samples, the training data contains 38 samples with 27 in ALL and 11 in AML and the rest 34 samples with 20 in ALL and 14 in AML compose the test data. Each predictor variable is standardized to have zero mean and unit variance (Dudoit et al., 2002; Fan and Fan, 2008). Then two sample t-tests against the two classes for each variable is performed such that variables with large t-tests statistics are ranked as significant variables. It has been shown that such a procedure can select all the important variables (Fan and Fan, 2008). Here I choose $p = 200$ and $p = 500$ top significant variables as the two settings for the predictors in the classification.

Table 2.4 reports the misclassification rates of the test data for the SLDA-ME, SLDA-EBIC, GLDA, Scout, FAIR and LPD. Similarly as in the simulation, I adopt 10-fold cross

Table 2.4: Misclassification rate of the proposed methods compared with other approaches for Leukemia data under the same selected genes

	SLDA-ME	SLDA-EBIC	GLDA	Scout	FAIR	LPD
$p = 200$	8.82%	11.76%	20.59%	11.76%	17.65%	17.65%
$p = 500$	8.82%	11.76%	20.59%	11.76%	17.65%	14.71%

validation for tuning parameters selection in SLDA-ME, Scout and LPD. The parameters $\gamma_1 = \gamma_2 = 0.5$ of EBIC criterion is used for SLDA-EBIC. The results in Table 2.4 shows consistent performance under $p = 200$ and $p = 500$. Specifically, SLDA-ME makes three misclassification errors out of the 34 test samples, slightly better than SLDA-EBIC which makes 4 errors. It appears that the performances of SLDA-ME and Scout are comparable, but better than LPD and FAIR. Clearly, the GLDA does not work very well with seven misclassification errors out of the 34 test samples.

To further evaluate the performance of the proposed methods, under the $p = 200$ setting, I merge the original training and test data sets into one data set. Then I randomly partition the merged data into two parts: half of the whole samples as a new training data set and the remaining half as a new test data set. The SLDA-ME, SLDA-EBIC, GLDA, Scout, FAIR and LPD are performed and the misclassification errors are calculated using the new test data. I repeat this procedure 50 times, and each time the misclassification errors are calculated for all the methods in comparison. Figure 2.1 displays boxplots of the misclassification rates for the six methods. The results in Figure 2.1 are consistent with the result under the original Leukemia data in Table 2.4. The proposed SLDA-ME works best, followed by SLDA-EBIC and FAIR. The LPD provides slightly larger errors while Scout and GLDA do not provide accurate classification.

2.5.2 Prostate Cancer Data

The Prostate Cancer data (Singh et al., 2002) is available at http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi. The training data contains 102 samples with 52 samples collected from tumor patients and the other 50 samples from normal people. The test data contains 34

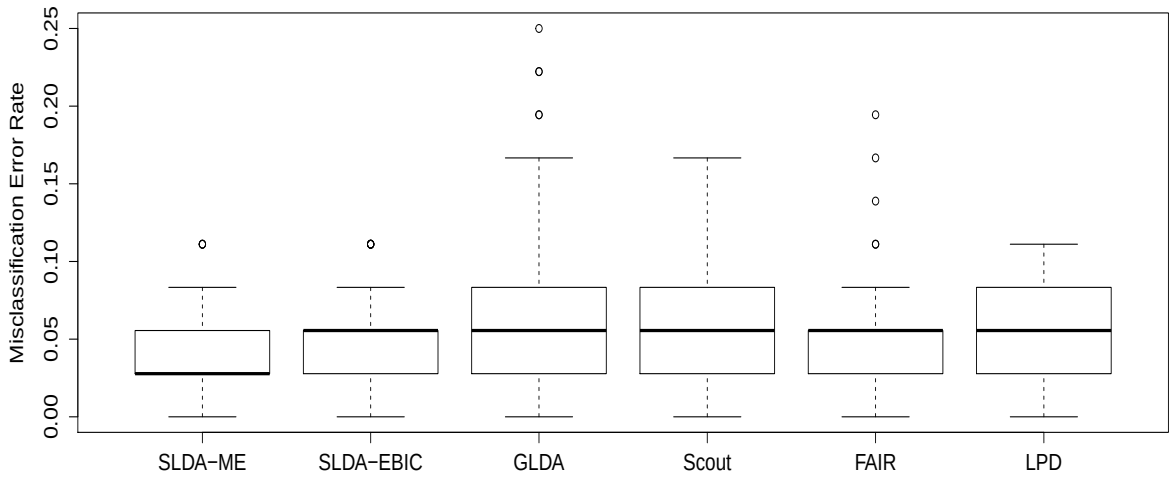


Figure 2.1: Misclassification error comparison for proposed methods with other approaches under the randomly splitting training and test data from Leukemia data under p=200

Table 2.5: Misclassification error of the proposed methods compared with other approaches for Prostate Cancer data under the same selected genes

	SLDA-ME	SLDA-EBIC	LDA	Scout	FAIR	LPD
$p = 200$	5	10	10	14	11	7
$p = 500$	5	10	11	14	11	8

samples, 25 of which come from tumor patients and the other 9 samples come from normal people. The same 12600 genes are measured for each patient. Following the same procedure as analyzing the Leukemia data, I perform the data standardization and variable screening on the Prostate Cancer data. Each variable is standardized by subtracting the mean and divided by the standard error. Then two sample t-tests are performed on each variable and p variables with the largest t statistics are chosen, where $p = 200$ or 500. I adopt 10-fold cross validation for tuning parameters selection in SLDA-ME, Scout and LPD. The parameters $\gamma_1 = \gamma_2 = 0.5$ of EBIC criterion is used in SLDA-EBIC. The results in Table 2.5 shows the misclassification rates for Prostate Cancer data under the proposed methods and other approaches.

From Table 2.5 we can see that SLDA-ME yields five misclassified cases out of 34 test samples, followed by LPD which makes seven errors when $p = 200$ and eight errors when

$p = 500$. Whereas the SLDA-EBIC does not perform very well and yields ten misclassified cases. So under this specific data, the ME criterion for SLDA is a more appropriate criterion than EBIC. The FAIR and GLDA perform comparable to the SLDA-EBIC. While Scout does not perform well on this data set.

To further examine the performance of the proposed methods, under the $p = 200$ setting, I merge the original training and test data into one data set and perform the random splitting procedure to form new training and test data. Half of the randomly selected data is treated as the new training data and the other half as the new test data. I repeat this procedure 50 times. Figure 2.2 displays the boxplots regarding the misclassification rate for each method. The results are consistent with the result from the original data displayed in Table 2.5. The proposed methods has a lower misclassification error thus the proposed methods performs better than the other methods compared.

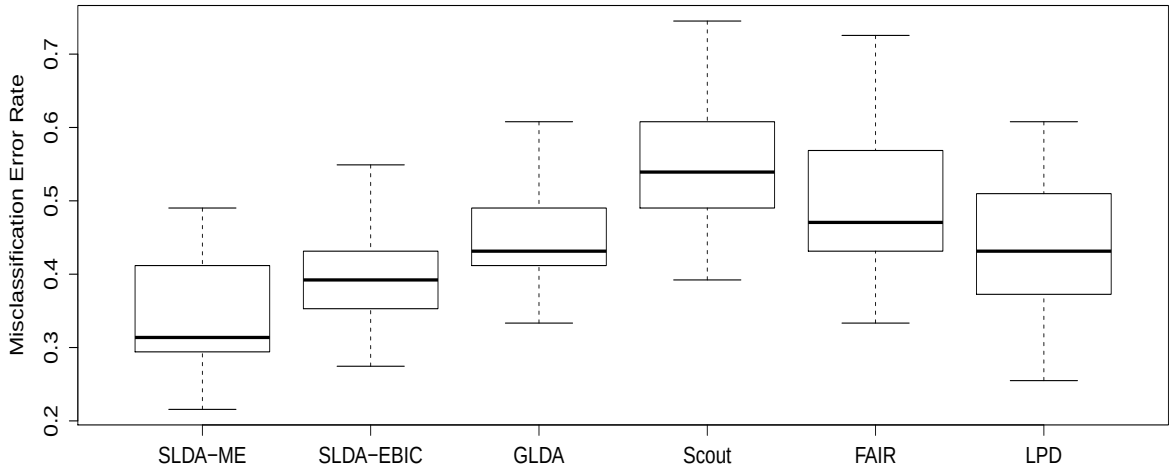


Figure 2.2: Misclassification error comparison for proposed methods with other approaches under the randomly splitting training and test data from Prostate Cancer data under p=200

2.6 Discussion

In this chapter a two-class classification method is proposed with separate regularization terms on the inverse covariance matrix $\boldsymbol{C}$ and the mean difference between the two classes $\boldsymbol{\delta}$ to maintain their sparse structures simultaneously. An iterative procedure consists of a graphical lasso step and a lasso step is proposed to estimate parameters $\boldsymbol{C}$ and $\boldsymbol{\delta}$. I have considered misclassification error and EBIC as the criteria to choose the tuning parameters. In terms of computational cost, SLDA-ME can be more expensive than SLDA-EBIC. The proposed method outperforms other methods in terms of classification under different simulation settings and in real data examples.

In this chapter, the proposed method focuses on the two-class classification problem. But the proposed method can also be extended for multi-class classification. When there are more than two classes, the objective function will no longer be in the same form as (2.4). Generally, when there are K classes, the second penalty term in (2.4) need to be changed. A straightforward idea is to add $K-1$ tuning parameter λ_{k+1} on $\boldsymbol{\delta}_k = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_k, k = 2, \cdots, K$ besides the tuning parameter λ_1 on $\boldsymbol{C}$. To reduce the computational burden, I will consider to use the contrast to combine the $K-1$ mean differences shrinkage terms into one. Instead of penalizing $\boldsymbol{\delta}_k, k = 1, \cdots, K-1$ separately, they can be incorporated into a single penalizing term on $\boldsymbol{A\mu}$, where $\boldsymbol{A}$ is a full rank contrast matrix, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1', \cdots, \boldsymbol{\mu}_K')'$ where $\boldsymbol{\mu}_i, i = 1, \cdots, K$ is the mean vector of each class. Following the same derivation in Section 2.2, the objective function will be in a similar form as the profile likelihood function (2.7) which can be iteratively solved using Algorithm 1 under minor modification.

As a classification procedure based on LDA, the proposed method assumes that the covariance matrix $\boldsymbol{\Sigma}$ remains the same across different classes. While sometimes this assumption may be violated. Traditionally a quadratic discriminant analysis (QDA) can be performed under different covariance structures assumption. QDA contains more parameters compared with LDA. Although there are regularized forms of QDA proposed (Friedman, 2008; Hastie, Tibshirani and Friedman, 2008), they may not perform very well when $p > n$. To extend

the proposed method in Section 2.2, we can consider adding one more penalization term on $||\boldsymbol{C}_1 - \boldsymbol{C}_2||_1$, where $\boldsymbol{C}_1 \triangleq \boldsymbol{\Sigma}_1^{-1}$ and $\boldsymbol{C}_2 \triangleq \boldsymbol{\Sigma}_2^{-1}$ are the inverse covariance matrices for the two classes. To solve the new optimization for parameter estimation, we can adopt Danaher et al. (2014) and change the Step 2 in Algorithm 1 accordingly. Other joint inverse covariance matrices estimation approaches can be found in Danaher et al. (2014) and Guo et al. (2011).

Chapter 3 A Two-stage Risk Model Building and Evaluation in Reject Inference

3.1 Introduction

Fraud is a common criminal activity existing in almost every business sector (Button et al., 2009). Fraudsters always intend to get commercial products and services by deception, causing economic loss for individuals and companies. With the prevalence of electronic commerce, more and more frauds occur in online purchases. To prevent frauds from online transactions, various fraud detection techniques are used to assess the incoming transactions to be rejected or accepted, and consequently identify fraudsters from normal customers (Ngai et al., 2010).

Scorecards (Bolton and Hand, 2002; Delamaire et al., 2009) are widely used in financial institutions for fraud check. As with scorecards, decision makers in online business usually build a risk model for fraud detection. The records of each online transaction contains rich information such as account number, IP address, payment instrument type and payment dollar amount. Generally, each transaction is assigned a score by the risk model based on its information and features. Transactions with high scores of being fraud are rejected. A transaction becomes a settled transaction when it also passes bank investigation. Because accepted transactions have negligible differences with settled transactions, I simply refer them as accepted transactions in my later sections. Figure 3.1 illustrates the fraud check system. Therefore, the quality of the risk model is important because the acceptance/rejection decision for online transaction will greatly affect the business profit.

Data on accepted and rejected transactions are available for constructing the risk model. Since the status of being fraud or not is easy to obtain for the accepted transactions, but not for the rejected ones, a straightforward way is to use solely the accepted transactions for constructing the risk model. However, the population of rejected transactions is different from

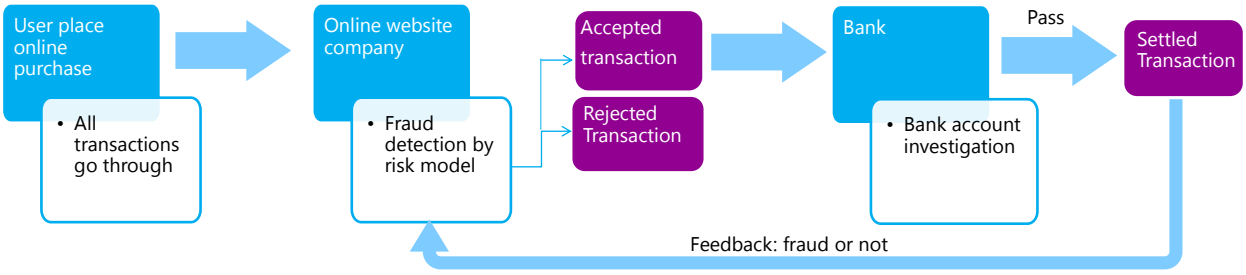


Figure 3.1: Flow chart of the fraud detection system

that of the accepted because of the selective nature of the rejection region (Copas and Li, 1993). Since the accepted transactions are not a good representative subset of the whole population of transactions, excluding rejected transactions data will result in bias in model estimation. To construct an accurate risk model, it calls for "reject inference" (Chen and Astebro, 2001; Feelders, 2003; Crook and Banasik, 2004), which incorporates both the accepted and rejected transactions appropriately.

There are several challenges in reject inference in order to accurately assess the fraud/non-fraud status for a transaction. For a rejected transaction, its status of being fraud or not is unknown. If all the rejected transactions are treated as fraud cases, the resultant model could mistakenly reject non-fraud transactions based on their risk scores of being fraud. If all the rejected transactions are removed from the data, the resultant model could fail to detect certain fraud patterns if many rejected transactions are fraud cases.

Another challenge in reject inference is how to incorporate the rejected transactions properly in the model evaluation criteria. Rejected transactions usually cause the difficulty of estimating two major parameters in model evaluation. The first one is the retrial rate due to the reject experience. Define a user as an online account owner who performs a purchase action. When one transaction is rejected, normally a user will insist on their intention by placing another purchase trial. It may not be appropriate to decide a transaction to be a retrial based on its purchase amount since a user may purchase a substitute item that is different with the original one. It is also difficult to choose a time interval between two trials to define a retrial. The second parameter is the percentage of normal/fraud transactions among

the rejected transactions. The A/B testing technique (Kohavi et al., 2012) is one method for estimating the two unknown parameters. In online business, the key idea of A/B testing is to allow a small percentage of randomly selected users to be 100% accepted, known as control group B. By flagging the transactions that should be rejected in group B, the retrial rate and percentage of normal/fraud transactions among the rejected transactions can be accurately estimated. However, the implementation of A/B testing is costly because the fraudsters in group B are not rejected during this procedure.

In this chapter, I focus on constructing a two-stage risk model and developing a novel model evaluation criterion for reject inference. As shown in Figure 3.1, transactions are first evaluated by the risk model to determine their acceptance/rejection status, then the fraud status of accepted transactions will not be known until they are settled by bank investigation. Since this fraud detection system has a natural hierarchical structure, I propose a two-stage modeling approach for fully exploiting the fraud pattern. The first-stage model focuses on capturing fraud patterns in the rejected transactions (*already captured fraud*) and fraud patterns in the accepted transactions (*missing fraud*) simultaneously. The second-stage model is designed to further capture those missing frauds. Moreover, a weighting scheme for different types of transactions is applied to emphasize their roles in the two-stage modeling, making the proposed method more effective to identify the missing fraud. The major advantage of this two-stage model is that the information of rejected transactions are properly used, such that the acceptance/rejection decision is made at both stages accordingly. To find an optimal two-stage model, I further developed a novel model evaluation criterion based on net profit value (NPV). By adjusting the calculation of NPV, the proposed criterion can mimic the performance of NPV directly calculated from group B. It therefore provides insight on how to accommodate the rejected transactions into the model evaluation criterion.

The remainder of this chapter is organized as follows. In Section 3.2, I give a brief review on existing reject inference techniques. The details of the general data structure are described in Section 3.3. In Section 3.4, the two-stage model is proposed for fraud detection. Section 3.5 details the development of new model evaluation criteria. In Section 3.6, I elaborate

the performance of the proposed method through a real case study of Microsoft Xbox online transactions data. Section 3.7 concludes this chapter with some discussion.

3.2 Literature Review

Reject inference (Chen and Astebro, 2001; Feelders, 2003; Crook and Banasik, 2004) has been widely studied in different applications. Under the context of online business, different reject inference methods can be generally grouped into three categories in terms of model assumptions. The first category assumes that the accepted transactions can represent the whole population of all transactions and contain the full information for risk modeling. This assumption is very strong and often too expensive to be satisfied. The A/B testing, mentioned in Section 3.1, is a known technique of this kind. If properly used, the A/B testing can lead to an accurate risk model. However, this technique is costly since all the fraud cases in group B are accepted during the A/B testing procedure.

The other two types of model assumptions are referred to as missing at random (MAR) and missing not at random (MNAR) (Feelders, 2003), respectively. A thorough discussion of the missing schemes in reject inference can be found in Shaun et. al. (2013). The MAR assumes that the fraud status of rejected transaction is missing at random (Feelders, 2003). For each transaction, denote by a its acceptance status with value 1 being accepted, and 0 otherwise. Similarly, define y as its fraud status of whether this transaction is fraud or not (labelled as 1 or 0, respectively). Then the assumption of MAR can be written as

$$P(y = 1|\boldsymbol{x}, a = 1) = P(y = 1|\boldsymbol{x}, a = 0),$$

where $\boldsymbol{x}$ is a vector of predictor variables of the transaction. In another word, the MAR assumes that the rejected samples and the accepted samples come from the same population for fraud case.

There are several methods proposed under the assumption of MAR. One commonly used method is called augmentation (Montrichard, 2008; Banasik and Crook, 2007). The corre-

sponding risk model is constructed by reweighing the accepted transactions. For example, Hsia (1978) assumed that the percentage of fraud among the accepted transactions is the same as that among the rejected transactions. This assumption is unrealistic in practice since it is reasonable to assume there are more fraud in the reject class. Another augmentation method (Montrichard, 2008) employs a two-step modeling process. It first models the acceptance/rejection status for all the transactions. Then a weight of $1/P(a = 1)$ is imposed on the accepted transactions for classification of being fraud or not. This method can result in large bias in model estimation due to large weight $1/P(a = 1)$ on some observations.

Another widely used approach under MAR is called extrapolation (Hand and Henley, 1993). In Siddiqi (2006), a classification model is constructed to classify the normal/fraud status by only using the accepted observations. Then the rejected observations are scored according to this model. Feelders (1999) assumed that the distributions of risk score for the rejected observations and accepted observations came from a particular family of distribution. Then the unknown fraud/non-fraud status for rejected observations are treated as missing data. However, the assumption of the method is questionable when the accepted and rejected population is not homogeneous.

Note that the assumption of MAR often does not hold in reality because of a selective reject nature. A more reasonable assumption is missing not at random (MNAR). The MNAR assumes that the distribution of rejected population is different from that of the accepted population:

$$P(y = 1|\boldsymbol{x}, a = 1) \neq P(y = 1|\boldsymbol{x}, a = 0).$$

Joanes and Derrick (1993) developed an iterative method under this framework. They constructed a prior probability of being fraud in the rejection region. Then they iteratively conducted the classification and modify the prior probability. Another example under MNAR is called two-stage bivariate probit model (Heckman, 1979; Copas and Li, 1997; Chen, 2001). This method separates the fraud/non-fraud status and the good/bad status into two models

and links them together under the bivariate normal error assumption:

$$a = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon_1$$

$$y = \boldsymbol{x}'\boldsymbol{\gamma} + \epsilon_2,$$

Note that the fraud status y is only observed when the acceptance status $a = 1$. The error terms ϵ_1 and ϵ_2 follow a bivariate normal distribution. However, the estimation in the bivariate probit model is not very robust and the multicollinearity problem often exists (Puhani, 2000).

My work is under the assumption of MNAR. I propose a two-stage method to build the risk model for reject inference. My two-stage model differs from Heckman′s two-stage bivariate probit model. The acceptance/rejection and the fraud/non-fraud information are used in both stages. The first stage model focuses on rejecting transactions with sufficiently high risk scores and the second stage model is designed to further capture missing fraud undetected from the first stage model.

3.3 Portfolio Decomposition

The whole portfolio is first decomposed into several subsets of transactions based on their acceptance/rejection and fraud/non-fraud status. Since each transaction is carried out by a specific account owner, the transactions conducted by the fraudster's account are considered to be fraud transactions. Therefore, it is important to consider the account information of transactions for portfolio decomposition. Moreover, when one transaction is rejected, the account owner will often make several retrials before successfully placing a purchase or giving up. Such retrial information will also be useful in the portfolio decomposition to improve the risk modeling and model evaluation.

Let us denote t to be a transaction and $a(t)$ to be the account owner of the transaction. Suppose the whole portfolio contains a set of n transactions as $S = \{t_1, \ldots, t_n\}$. For notation convenience, I denote being fraud as being bad and being non-fraud as being good. Thus, the

S can be partitioned into three mutually exclusive and exhaustive subsets as $S_{re},\ S_{ag}$, and S_{ab}, representing rejected, accepted good and accepted bad transactions, respectively. By defining $d(t_i)$ to be the fraud response of t_i, we can assign $d(t_i) = 1, 0, \mu$ to represent a transaction to be bad, good and rejected respectively:

$$d(t_i) = \begin{cases} 1, & \text{if } t_i \in S_{ab} \\ 0, & \text{if } t_i \in S_{ag} \\ \mu, & \text{if } t_i \in S_{re}. \end{cases}$$

Note that the partition of $S = S_{re} \cup S_{ag} \cup S_{ab}$ does not take into consideration of account information of transactions. Although the fraud status of rejected transaction is generally unknown, we can still figure out the fraud status of certain transactions by using account information. Normally the rejected transactions are retrials before their purchase intentions are accepted. So the true fraud status of those rejected transactions can be traced from the corresponding accepted transactions under the same account if there are any. Hence, we can partition the whole transaction set S into a few non-overlapping subsets as follows. For each transaction t_i, define the index set $I_i = \{j : a(t_j) = a(t_i)\}$ to include the indices of transactions with the same account owner as t_i. Then I denote $S_{RG} = \{t_i \in S : \exists h \in I_i, d(t_h) = \mu, \text{ and } \exists l \in I_i, d(t_l) = 0\}$ as the subset of transactions under rejected good (RG) users. It means that S_{RG} is the set of transactions whose account owners have, at least, one accepted transaction and one rejected transaction in the data. Similarly, I define $S_{RB} = \{t_i \in S : \exists h \in I_i, d(t_h) = \mu, \text{ and } \exists l \in I_i, d(t_l) = 1\}$ as the subset of transactions under rejected bad (RB) users. Note that few users may have both accepted good and bad transactions. In this work, I simply ignore such users since they rarely occur. I also define $S_{RU} = \{t_i : \forall h \in I_i, d(t_h) = \mu\}$ as the set of transactions of reject unknown (RU) users. Denote by $S_{AG} = \{t_i : d(t_i) = 0 \text{ and } \forall h \in I_i, d(t_h) \neq \mu\}$ the subsets of transactions from accepted good (AG) users and $S_{AB} = \{t_i : d(t_i) = 1 \text{ and } \forall h \in I_i, d(t_h) \neq \mu\}$ the accepted bad (AB) users.

The so-called missing fraud mentioned in Section 3.1 refers to transactions in S_{AB}. Clearly, we can see that $S = S_{RG} \cup S_{RB} \cup S_{RU} \cup S_{AG} \cup S_{AB}$. Following such a decomposition of transactions, I define $\tilde{d}(t_i)$ to be the adjusted response:

$$\tilde{d}(t_i) = \begin{cases} 1, & \text{if } t_i \in S_{AB} \cup S_{RB} \\ 0, & \text{if } t_i \in S_{AG} \cup S_{RG} \\ \tilde{\mu}, & \text{if } t_i \in S_{RU}. \end{cases} \tag{3.1}$$

The decomposition of transaction data in (3.1) will be used throughout this chapter.

3.4 Reject Inference Methodology

Reject inference here is generally a classification problem of distinguishing fraud transactions from normal transactions with certain non-randomly selected rejected observations. As stated in Section 3.1, there are rich information collected for each transaction t_i, which can be used as independent variables to predict its fraud/non-fraud status. Denote them as $\boldsymbol{x}_i = \{x_{i1}, \cdots, x_{ip}\}, i = 1, \cdots, n$.

I first screen the data for the denoising purpose. Whenever an online transaction takes place, it involves with a certain purchase amount. Let X_1 be the variable representing the purchase amount. Using X_1 for thresholding, I accept the transactions t_i's if $x_{i1} < u$ where u is a threshold value. Such a screening procedure is to avoid possible small scale noise. Then the rest of the transactions will be handled by the risk model. There are two reasons behind this step. Firstly, most fraudsters tend to acquire merchandise with large purchase amount. Secondly, even though there exist frauds with small purchase amount, the percentage and financial loss are rather small.

3.4.1 Two-Stage Modeling

I start with the key ideas of the proposed two-stage model. The details of each stage will be described in Section 3.4.1.1 and 3.4.1.2, respectively.

Let us define $R \in \{0, 1\}$ to be the rejection response with $R = 0$ being accepted and 1 being rejected. Denote $Y \in \{0, 1\}$ to be the fraud response with $Y = 1$ being fraud, which is only observed when a transaction is in $S_{RB} \cup S_{RG} \cup S_{AB} \cup S_{AG}$. For $\{t_i : t_i \in S_{RU}\}$, the fraud response R is missing. When $R = 1$, the transaction is rejected and the corresponding fraud response Y is not observed. The objective of the risk modeling is to model $P(Y = 1|\boldsymbol{x})$, where $\boldsymbol{x}$ is a realization of the set of independent variables $\{X_1, X_2, \cdots, X_p\}$.

Note that there exists certain association between Y and R. As stated in (3.1), for any transaction, its decision outcome is either 1, 0 or $\tilde{\mu}$. To take into account for the association between Y and R, I consider the following model structure:

$$\begin{aligned} P(Y = 0, R = 0|\boldsymbol{x}) &= 1 - f_1(\boldsymbol{x}), \\ P(R = 1|\boldsymbol{x}) &= f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})P(R = 0|\boldsymbol{x}), \\ P(Y = 1, R = 0|\boldsymbol{x}) &= f_2(\boldsymbol{x})P(R = 0|\boldsymbol{x}), \end{aligned}$$

Clearly, one can see that $P(Y = 0, R = 0|\boldsymbol{x}) + P(Y = 0, R = 1|\boldsymbol{x}) + P(R = 1|\boldsymbol{x}) = 1$. We can also obtain that $P(R = 0|\boldsymbol{x}) = (1 - f_1(\boldsymbol{x}))/(1 - f_2(\boldsymbol{x}))$. Therefore, we have

$$\begin{aligned} P(Y = 1|\boldsymbol{x}) &= P(Y = 1|R = 0, \boldsymbol{x})P(R = 0|\boldsymbol{x}) + P(Y = 1|R = 1, \boldsymbol{x})P(R = 1|\boldsymbol{x}) \\ &= \frac{f_2(\boldsymbol{x})(1 - f_1(\boldsymbol{x}))}{1 - f_2(\boldsymbol{x})} + P(Y = 1|R = 1, \boldsymbol{x})\frac{f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})}{1 - f_2(\boldsymbol{x})}. \end{aligned} \tag{3.2}$$

However $P(Y = 1|R = 1, \boldsymbol{x})$ is unknown in the reject inference. By defining $\alpha(\boldsymbol{x}) \triangleq$

$P(Y=1|R=1,\boldsymbol{x})$, we can have an approximated model:

$$\begin{aligned} P(Y=1|\boldsymbol{x}) &= \frac{f_2(\boldsymbol{x})(1-f_1(\boldsymbol{x}))}{1-f_2(\boldsymbol{x})} + \alpha(\boldsymbol{x})\frac{f_1(\boldsymbol{x})-f_2(\boldsymbol{x})}{1-f_2(\boldsymbol{x})} \\ &= \alpha(\boldsymbol{x})f_1(\boldsymbol{x}) + (1-\alpha(\boldsymbol{x}))P(Y=1,R=0|\boldsymbol{x}) \\ &\approx \alpha(\boldsymbol{x})f_1(\boldsymbol{x}). \end{aligned} \tag{3.3}$$

The approximated model (3.3) is based on $P(Y=1|\boldsymbol{x}) = P(Y=1,R=1|\boldsymbol{x}) + P(Y=1,R=0|\boldsymbol{x})$. Here the approximation is provided by the perception that the fraud rate of the rejected transaction is much larger than the fraud rate of the accepted transaction, i.e., $P(Y=1,R=1|\boldsymbol{x}) >> P(Y=1,R=0|\boldsymbol{x})$. Note that we would expect a large proportion of rejected transactions to be bad and the majority of accepted transactions to be good. It means that both $P(Y=1,R=0|\boldsymbol{x})$ and $(1-\alpha(\boldsymbol{x}))$ are small, which leads to the approximation in (3.3).

By using the score of $\alpha(\boldsymbol{x})f_1(\boldsymbol{x})$ as a reasonable approximation to $P(Y=1|\boldsymbol{x})$, one may be able to identify fraud or non-fraud. However, $\alpha(\boldsymbol{x})$ is unknown and we can only build score from $f_1(\boldsymbol{x})$. Since $\alpha(\boldsymbol{x})f_1(\boldsymbol{x})$ will shrink the value of $f_1(\boldsymbol{x})$, using $f_1(\boldsymbol{x})$ alone cannot help find out some of the fraud especially when $\alpha(\boldsymbol{x})$ is small. The score $f_1(\boldsymbol{x})$ is only useful to identify transactions which has high likelihood to be suspected fraud or evident non-fraud. Denote the two threshold values by τ_1 and τ_2 respectively. The estimated fraud response $\hat{y}=1$ if $f_1(\boldsymbol{x}) \geq \tau_1$, $\hat{y}=0$ if $f_1(\boldsymbol{x}) < \tau_2$. For the transaction with very large score of $f_1(\boldsymbol{x})$, I consider it as a fraud case. In contrast, transactions with very low scores of $f_1(\boldsymbol{x})$ are considered to be non-fraud cases. A simple derivation leads to

$$f_1(\boldsymbol{x}) = P(R=1|\boldsymbol{x}) + P(Y=1,R=0|\boldsymbol{x}). \tag{3.4}$$

That is, estimating $f_1(\boldsymbol{x})$ is to classify the transactions in $R=1$ and $Y=1,R=0$ as one class and the remaining as the other class. It implies that the fraud response of transactions in S_{RU} are treated the same as those in $S_{RB} \cup S_{AB}$. This perception comes from the nature

of the rejected transactions. The transactions in S_{RU} are suspected to be bad and thus are rejected by the online decision system. It is believed that a lot of fraud cases are captured in the rejected transactions. In order not to miss those already captured fraud, it is helpful to label transactions in S_{RU} same as known fraud transactions when constructing the risk model.

After removing the highly evident transactions, a second stage model is needed to quantify the fraud status of the remaining data. For the remaining transactions, it is no longer that easy to distinguish fraud in the rejected transactions from fraud in the accepted transactions. Therefore, one reasonable approximation is $P(Y = 1|R = 1, \boldsymbol{x}) \approx P(Y = 1|R = 0, \boldsymbol{x})$. Based on (3.2), we have the approximate model:

$$\begin{aligned} P(Y = 1|\boldsymbol{x}) &\approx \frac{f_2(\boldsymbol{x})(1 - f_1(\boldsymbol{x}))}{1 - f_2(\boldsymbol{x})} + P(Y = 1|R = 0, \boldsymbol{x})\frac{f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})}{1 - f_2(\boldsymbol{x})} \\ &= f_2(\boldsymbol{x}). \end{aligned}$$

The second stage model $P(Y = 1|\boldsymbol{x}) \approx f_2(\boldsymbol{x})$ attempts to identify the fraud undetected in stage one. Denoting by τ_3 another threshold value, then the estimated fraud response $\hat{y} = 1$ if $f_2(\boldsymbol{x}) \geq \tau_3$. It is easy to see that $f_2(\mathbf{x}) = P(Y = 1|R = 0, \boldsymbol{x})$.

Combining the above two stages, I propose to quantify the fraud response by approximating $P(Y = 1|\boldsymbol{x})$ as:

$$P(Y = 1|\boldsymbol{x}) \approx I(f_1(\boldsymbol{x}) > \tau_1 \cup f_1(\boldsymbol{x}) < \tau_2)f_1(\boldsymbol{x}) + I(\tau_2 < f_1(\boldsymbol{x}) < \tau_1)f_2(\boldsymbol{x}). \tag{3.5}$$

Clearly, the two-stage model makes acceptance/rejection decision using both scores $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$. The following two sections will detail the estimation of the two scores.

3.4.1.1 Stage-One Modeling

The goal of stage-one modeling is to detect the already captured fraud patterns in rejected transactions as well as some missing frauds in S_{AB}. To estimate $f_1(\boldsymbol{x})$ from the data based

on (3.4), I consider the classification problem with the label of $\tilde{y}(t_i)$ as follows:

$$\tilde{y}(t_i) = \begin{cases} 1, & \text{if } t_i \in S_{RB} \cup S_{AB} \cup S_{RU} \\ 0, & \text{if } t_i \in S_{RG} \cup S_{AG}. \end{cases} \tag{3.6}$$

It means that transactions that are either fraud or rejected with unknown fraud status are coded as 1 and the rest transactions coded as 0. Although there are various classification methods available in literature (Hastie, Tibshirani and Friedman, 2001; Kotsiantis, Zaharakis and Pintelas, 2006), I adopt the logistic regression for its ease of interpretation and fast computation. Thus, the estimation of $f_1(\boldsymbol{x})$ can be obtained from

$$f_1(\boldsymbol{x}_i) = \Pr(\tilde{y}(t_i) = 1|\boldsymbol{x}_i) = \frac{1}{1 + e^{-(\beta_0^{(1)} + \Sigma_{j=1}^{p} \beta_j^{(1)} x_{ij})}}, \tag{3.7}$$

where $\{\beta_0^{(1)}, \cdots, \beta_p^{(1)}\}$ are parameters.

To estimate the parameters efficiently, I propose a weighted logistic regression with weight w ($w \geq 1$) on transactions $\{t_i : t_i \in S_{AB}\}$ and weight 1 on all other transactions. There are several reasons of using such a weighting scheme. First, a good risk model needs to detect the missing fraud. By imposing a large weight on the observed bad transactions, importance of missing fraud are emphasized. Second, because the data is highly unbalanced with usually only less than 5% fraudulent transactions, a large weight on bad transactions would balance the data to some extent. It is known that the number of non-fraud transactions is much larger than the number of fraud transactions. For example, the real data used in Section 3.6 has about 99% of the accepted transactions being good transactions. Finally, I do not give more weight to the rejected transactions although we also want to catch the already captured frauds in the rejection region. The fraud pattern in the rejected transactions is usually more homogeneous compared with that of the missing fraud. While missing fraud is difficult to detect due to variation in fraud types and similarities with good transactions. Hence transactions in S_{RB} is not assigned a larger weight.

Under the weighted logistic regression, I estimate the parameters $\boldsymbol{\beta}^{(1)} = (\beta_0^{(1)}, \cdots, \beta_p^{(1)})'$ by maximizing the following log-likelihood function:

$$L_1(\boldsymbol{\beta}^{(1)}) = \sum_{t_i \in S_{AB}} z \times [(\tilde{y}(t_i) - 1) \times (\beta_0^{(1)} + \Sigma_{j=1}^{p} \beta_j^{(1)} x_{ij}) - \log(1 + e^{-(\beta_0^{(1)} + \Sigma_{j=1}^{p} \beta_j^{(1)} x_{ij})})] + \sum_{t_i \in S_1 / S_{AB}} [(\tilde{y}(t_i) - 1) \times (\beta_0^{(1)} + \Sigma_{j=1}^{p} \beta_j^{(1)} x_{ij}) - \log(1 + e^{-(\beta_0^{(1)} + \Sigma_{j=1}^{p} \beta_j^{(1)} x_{ij})})],$$

where $S_1 = \{t_i \in S : x_{i1} \geq u\}$ is the set of transactions after the pre-thresholding step. Once $f_1(\boldsymbol{x})$ is estimated, every transaction can get a score as $\hat{f}_{1i}, i = 1, \cdots, n$. If $\hat{f}_{1i} \geq \tau_1$, transaction t_i is rejected; if $\hat{f}_{1i} < \tau_2$, t_i is accepted; otherwise, we use stage-two modeling.

Note that the stage-one modeling uses all the transactions including the rejected transactions. Thus it can incorporate both the missing fraud and already captured fraud patterns. If the rejected transactions are removed in model estimation, we will lose the information of fraud already captured in the rejection region.

3.4.1.2 Stage-Two Modeling

Using stage-one modeling, the fraud status can be estimated for transactions with high score $\hat{f}_{1i} \geq \tau_1$ or low score $\hat{f}_{1i} \leq \tau_2$. For the remaining transactions, the stage-two modeling of $f_2(\boldsymbol{x})$ is used to predict their fraud status. Specifically, I consider the logistic regression to estimate $f_2(\boldsymbol{x}) = P(Y = 1 | R = 0, \boldsymbol{x})$ using transactions

$$S_2 = \{t_i \in S_{AG} \cup S_{AB} : \tau_2 \leq \hat{f}_{1i} < \tau_1\}.$$

Here the transaction in S_{AB} is labelled as 1 and those in S_{AG} is labelled as 0. Then the parameters $\boldsymbol{\beta}^{(2)} = (\beta_0^{(2)}, \cdots, \beta_p^{(2)})'$ can be estimated by maximizing the log-likelihood function $L_2(\boldsymbol{\beta}^{(2)})$ where

$$L_2(\boldsymbol{\beta}^{(2)}) = \sum_{t_i \in S_2} [(\tilde{y}(t_i) - 1) \times (\beta_0^{(2)} + \Sigma_{j=1}^{p} \beta_j^{(2)} x_{ij}) - \log(1 + e^{-(\beta_0^{(2)} + \Sigma_{j=1}^{p} \beta_j^{(1)} x_{ij})})].$$

Note that in this stage-two modeling, I have excluded the data $\{t_i : t_i \in S_{RU} \cup S_{RG} \cup S_{RB}\}$. It is to make the modeling focus on the accepted transactions, and identify more missing fraud not captured in stage-one modeling. More importantly, as highly suspected fraud in the rejection region has been detected in stage-one model, the fraud/non-fraud status for rejected transactions with score $\tau_1 \leq \hat{f}_{1i} \leq \tau_2$ is more uncertain. It is also worth pointing out that this stage does not impose any weighting scheme for parameter estimation. It can keep the model estimation simple and stable. Moreover, the fraud pattern in the missing fraud is not homogeneous thus placing more weight on missing fraud will reduce the prediction power of a risk model.

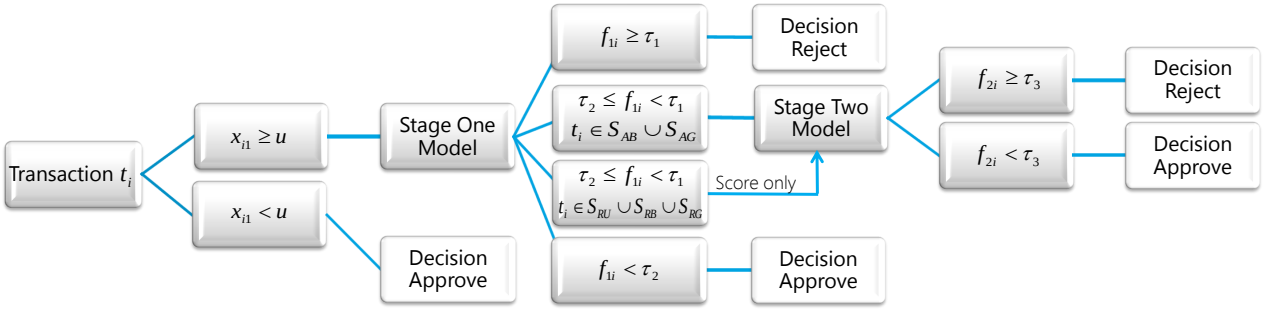


Figure 3.2: Flow chart of the proposed two-stage modeling procedure

Once obtaining the estimation of parameters, we can assign the score of $f_2(x)$ for all the remaining transactions $S_2 = \{t_i \in S_{AG} \cup S_{AB} : \tau_2 \leq \hat{f}_{1i} < \tau_1\}$ and $S_3 = \{t_i \in S_{RU} \cup S_{RB} \cup S_{RG} : \tau_2 \leq \hat{f}_{1i} < \tau_1\}$. Denote $\hat{f}_{2i}$ as the estimated value of f_2 for transaction t_i in stage-two modeling. If $\hat{f}_{2i} \geq \tau_3$, it is rejected, otherwise it is accepted.

Figure 2 shows a flow chart of the proposed two-stage modeling procedure. For any new transaction t^*, if its purchase amount is larger than u, it will get a score of f_1^* using the first stage model, otherwise it will simply get accepted. If the value $f_1^* \geq \tau_1$, it will be rejected; if $f_1^* < \tau_2$, it will be accepted. Otherwise, it will go through the second stage model and have a score of f_2^*. Then the fraud status of this transaction will be estimated by whether $f_2^* \geq \tau_3$ or not.

3.4.2 Choice of Tuning Parameters

The proposed two-stage modeling involves five tuning parameters: the purchase amount cut-off value u, the weight z in stage-one modeling and the threshold values τ_1, τ_2, τ_3. The cut-off value u is usually suggested subjectively by the domain knowledge. The rejection rate is often fixed at a desired level. Thus, once the threshold values τ_1 and τ_2 are determined, the threshold value τ_3 in stage-two modeling is also determined to maintain the fixed rejection rate. Therefore, the unknown tuning parameters are z, τ_1, τ_2. In practice I will use percentages p_1 and p_2 to replace threshold values τ_1 and τ_2. Let α be the fixed rejection rate. It means that in stage-one modeling, $p_1\alpha \times 100\%$ of transactions with the highest f_1 score are rejected and $p_2 \times 100\%$ of transactions with the lowest f_1 score are accepted. Clearly p_1 and p_2 have one-to-one mapping to τ_1 and τ_2. Denote the model evaluation criterion as $\Delta = g(z, p_1, p_2)$ given an estimated model. Using the validation data or testing data, I propose to choose the tuning parameter by using a greed procedure as follows,

Algorithm 2

Step 0: *Give the initial values of* $\hat{p}_1$ *and* $\hat{p}_2$.

Step 1: *Fix* $p_1 = \hat{p_1}$ *and* $p_2 = \hat{p_2}$ *values, solve the optimization* $\hat{z} = \arg\max_z g(z, p_1, p_2)$.

Step 2: *Fix the weight* $z = \hat{z}$, *solve the optimization* $\max_{p_1,p_2} g(z, p_1, p_2)$.

Step 3: *Go back to* Step 1 *and repeat* Step 1 *and* Step 2 *until* $\hat{z}, \hat{p}_1, \hat{p}_2$ *converge.*

Note that when the model evaluation criterion $g(z, p_1, p_2)$ is a non-convex function, tuning parameters selected by this method may fall in some local maximum point (Cances, Ehrlacher and Lelievre, 2011). Under my problem's context, starting from a reasonable initial setting of z, p_1, p_2, it appears that the proposed method provides a promising result. And the algorithm converges fast in the case study in Section 3.6.

3.5 Model Evaluation Criteria

For the reject inference in online business, the major objective is to maximize the portfolio net profit value (NPV). Therefore, the NPV is used as the model evaluation criterion in this work. However, the calculation of the NPV involves two issues in the rejection region: the retrial transactions and the mistakenly rejected transactions. Retrial transactions commonly occur due to rejected purchase intention. For the users with mistakenly rejected transactions, I refer them as the churn users. In this work, I will adjust the retrial transactions and churn users in the NPV calculation. The purpose of the adjustment on the validation data is to mimic the performance of NPV calculated from group B data such that the proposed adjusted NPV can be used as alternative to A/B testing. In addition, I will also adjust the calculation of the number of false positives (FP) and the number of false negatives (FN) as the evaluation of classification accuracy.

3.5.1 The Formulation of NPV

In reality, only part of the revenue will turn into profit and this part is usually referred to as margin (Tinsley and Stetz, 2004). Denote m to be the margin portion of revenue, and let $v(t_i)$ be the purchase amount of transaction t_i. Based on the online decision system, the NPV of transaction t_i can be expressed as

$$\delta_0(t_i) = \begin{cases} m \times v(t_i), & \text{if } d(t_i) = 0, \\ -(1-m) \times v(t_i), & \text{if } d(t_i) = 1, \\ 0, & \text{if } d(t_i) = \mu. \end{cases} \tag{3.8}$$

Then the total NPV of all transactions can be calculated by $\Delta_0 = \Sigma_{i=1}^{n} \delta_0(t_i)$.

For a new model M used to make acceptance/rejection decision, denote by $\hat{y}(t_i)$ the decision

of transaction t_i from model M. Then $\hat{y}(t_i)$ has two status:

$$\hat{y}(t_i) = \begin{cases} 0, & \text{if } t_i \text{ is accepted under } M; \\ 1, & \text{if } t_i \text{ is rejected under } M. \end{cases}$$

Note that for a rejected transaction $t_i \in \{S_{RG} \cup S_{RB} \cup S_{RU}\}$, its value of $\hat{y}(t_i)$ is either 1 or 0 under model M. The real fraud status of $t_i \in S_{RU}$ is unknown so it is difficult to tell if a transaction $t_i \in S_{RU}$ is accepted or rejected rightly under this new model. The acceptance/rejection decision will then directly affect the action of gaining/losing profit when calculating NPV. The unknown fraud status will make the formulation of NPV $\delta_M(t_i)$ more complicated. Moreover, the retrials in the rejected transactions should only be counted once when calculating NPV. In many businesses it is very difficult to generate a retrial flag since there are various commodities type and human behaviour. To address these issues, several discount numbers are introduced to handle possible retrials and unknown fraud/non-fraud status.

Normally fraudsters' accounts do not contain normal transactions and normal users' accounts do not contain fraud transactions. An online account that contains both normal transaction and fraud transaction rarely occurs thus they are not considered in my study. Then the discount rates are defined based on this reasoning. For $t_i \in S_{RG}$, discount rate $d_{RG}(t_i)$ is used to reflect $t_i \in S_{RG}$ being a retrial, i.e.,

$$d_{RG}(t_i) = \begin{cases} 1, & \text{if } \exists t_j \in I_i, \hat{y}(t_j) = 0 \text{ and } d(t_i) = 0; \\ 0, & \text{otherwise.} \end{cases} \tag{3.9}$$

The status of retrials are decided by the fraud status of the accepted transaction under a certain user. The adjustment then adjust all the trials happened before the accepted transaction. For example, an user performed three transactions with the first two transactions rejected and

the third one accepted, the three trials will only be counted once when at least one of the three transactions is accepted under model M.

Similarly, for $t_i \in S_{RB}$, I use the discount rate $d_{RB}(t_i)$ to reflect $t_i \in S_{RB}$ being a retrial,

$$d_{RB}(t_i) = \begin{cases} 1, & \text{if } \exists t_j \in I_i, \hat{y}(t_j) = 1 \text{ and } d(t_i) = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{3.10}$$

For $t_i \in S_{RU}$, it will be more complicated since there may have mistakenly rejected transactions. Denote by S_{RU}^G and S_{RU}^B the good and bad transactions in S_{RU}, respectively. That is $S_{RU} = S_{RU}^G \cup S_{RU}^B$. To properly calculate $\{\delta_M(t_i),\ \text{if } \hat{y}(t_i) = 0, t_i \in S_{RU}\}$, I define r and s to be discount rates for non-fraud/fraud transactions in S_{RU} respectively, i.e.,

$$\begin{aligned} r &= d_{RU} \times \gamma; \\ s &= d_{RU} \times (1 - \gamma), \end{aligned} \tag{3.11}$$

where $\gamma = Pr(t \in S_{RU}^G | t \in S_{RU})$ is the proportion of good transactions in S_{RU}, and d_{RU} is the discount rate for $t_i \in S_{RU}$ being a retrial. It means two discount rates r and s are used to take into account of both unknown fraud status and the retrial problem. Note that d_{RU} equals 1 minus the retrial rate and it remains a constant for all $t_i \in S_{RU}$. An overall discount number d_{RU} is used due to lack of information on which specific transaction is retrial. Because of unknown fraud status, the average purchase amount of $t_i \in S_{RU}^G$ and $t_i \in S_{RU}^B$, $\bar{v}_{RU}^G$ and $\bar{v}_{RU}^B$ are used instead of individual $v(t_i)$ in calculating $\{\delta_M(t_i), t_i \in S_{RU}\}$. Hence, under model M,

the formulation of adjusted NPV $\delta_M(t_i)$ becomes:

$$\delta_M(t_i) = \begin{cases} m \times v(t_i), & \text{if } \hat{y}(t_i) = 0, t_i \in S_{AG}; \\ -(1-m) \times v(t_i), & \text{if } \hat{y}(t_i) = 0, t_i \in S_{AB}; \\ m \times d_{RG}(t_i) \times v(t_i), & \text{if } \hat{y}(t_i) = 0, t_i \in S_{RG}; \\ -(1-m) \times d_{RB}(t_i) \times v(t_i), & \text{if } \hat{y}(t_i) = 0, t_i \in S_{RB}; \\ r \times m \times \bar{v}^G_{RU} - s \times (1-m) \times \bar{v}^B_{RU}, & \text{if } \hat{y}(t_i) = 0, t_i \in S_{RU}; \\ 0, & \text{if } \hat{y}(t_i) = 1. \end{cases} \tag{3.12}$$

Then the total adjusted NPV is $\Delta_M = \sum_{i=1}^{n} \delta_M(t_i)$. Note that the value of $\bar{v}^G_{RU}$ and $\bar{v}^B_{RU}$ in (3.12) can come from business knowledge or analysis of current data, which may not be very accurate. From (3.12), one can see that even if the estimates for $\bar{v}^G_{RU}$ and $\bar{v}^B_{RU}$ are not very accurate, a good estimate of r and s can be used to calibrate the inaccuracy. Note that the adjusted NPV can be viewed as a linear function of r and s, i.e., $\Delta_M = \Delta_M(r, s)$. I will illustrate how to estimate these two values in Section 3.6.

3.5.2 The Formulation of False Positive and False Negative

Similar to adjusting the calculation of NPV in Section 3.5.1, I also make adjustment when calculating the number of false positives (FP) and false negatives (FN) such that the retrials before a successful trial will not be counted multiple times. Under model M, the adjusted FP

of each transaction is calculated as follows:

$$\mathrm{FP}(t_i) = \begin{cases} 1, & \text{if } \hat{y}(t_i) = 1, t_i \in S_{AG}; \\ d_{RG}(t_i), & \text{if } \hat{y}(t_i) = 1, t_i \in S_{RG}; \\ r, & \text{if } \hat{y}(t_i) = 1, t_i \in S_{RU}; \\ 0, & \text{otherwise .} \end{cases}$$

Then total number of false positives is $\mathrm{FP} = \sum_{i=1}^{n} \mathrm{FP}(t_i)$. Similarly, under model M, the calculation of adjusted FN can be as follows:

$$\mathrm{FN}(t_i) = \begin{cases} 1, & \text{if } \hat{y}(t_i) = 0, t_i \in S_{AB}; \\ d_{RB}(t_i), & \text{if } \hat{y}(t_i) = 0, t_i \in S_{RB}; \\ s, & \text{if } \hat{y}(t_i) = 0, t_i \in S_{RU}; \\ 0, & \text{otherwise .} \end{cases}$$

Then the total number of false negatives is written as $\mathrm{FN} = \sum_{i=1}^{n} \mathrm{FN}(t_i)$.

3.6 Case Study

In this case study, I will evaluate the performance of the proposed two-stage model in Section 3.4 and the adjusted model evaluation criteria in Section 3.5.

3.6.1 Data Description

The data set from Microsoft Xbox.com online purchase is used to illustrate the proposed methods. This website sells various accessories related to Xbox including games, token, game points, etc.. For such online purchase data, transactions that are suspected to be fraud are

rejected. It calls for the reject inference for data analysis. The A/B testing was applied for this data: the group B is composed from 10% randomly selected users being 100% accepted and the group A is composed from the rest 90% users being treated as regular business going through the online decision system. The total number of transactions is more than 300,000. Due to confidential reasons, there are certain figures not revealed here and so is the time window for this data. Specific information of the variables that are selected in the two-stage model is also masked for their sensitive nature to fraudsters. In group B data, all transactions are accepted and their fraud responses have been collected. Although there is no rejection region in group B, I have flagged transactions that are supposed to be rejected by the online decision system. All categorical variables are transformed into continuous variables (Anderson, 2007) and 56 variables are selected in total as independent variables based on business knowledge.

3.6.2 Comparison Results

I divide group A data into a training data set containing 70% randomly selected users' transactions and a validation data set containing the rest 30% users' transactions. The training data is used to estimate the two-stage risk model, while the validation and group B data are used for the calculation of NPV, model selection and result comparison. Since the calculation of adjusted NPV in (3.12) requires the estimate of $\bar{v}_{RU}^G$, $\bar{v}_{RU}^B$, r and s, I consider a heuristic estimation approach as follows.

When calculating the average purchasing amounts $\bar{v}_{RU}^G$ in S_{RU}^G and $\bar{v}_{RU}^B$ in S_{RU}^B, the challenges come from the retrials and unknown fraud status in S_{RU}. Let v_{RU}^G and v_{RU}^B be the total purchase amounts in S_{RU}^G and S_{RU}^B, respectively. I estimate v_{RU}^G and v_{RU}^B by discounting the purchase amount in group A, see Appendix C for details. By using the information that the proportion of good users in group A should be the same as that in group B, I can also properly estimate n_{RU}^G and n_{RU}^B, the number of good and bad transactions in S_{RU}, see Appendix D for details. Therefore, I can estimate $\bar{v}_{RU}^G$ and $\bar{v}_{RU}^B$.

To calculate r and s in (3.11), it needs the estimation of the discount rate d_{RU} and the probability $\gamma = Pr(t \in S_{RU}^G | t \in S_{RU})$. Although group B has 10% users and group A has

90% users, the number of transactions in group A is more than nine times of the transactions in group B. It is because many retrials exist in S_{RU} of group A data. Denote by n_c the total number of transactions in group B and n_s the total number of transactions of group A. Let n_{RU} be the total number of transactions in S_{RU}. Then d_{RU} can be estimated by $(9 \times n_c - n_s)/n_{RU}$. By getting the number of good transactions in group B with rejected flag, denoted as n_{aG}, the value of γ can be estimated by $\gamma = n_{aG}/n_c$.

For comparison with the proposed adjusted NPV, I also calculate the NPV under group B data. The calculation is straightforward since there is no rejected transaction. I also calculate the *unadjusted NPV* under the validation data, which ignores the retrials in $S_{RG} \cup S_{RB}$. That is, transactions in S_{RU} are simply treated as bad transactions because of the reject nature. Denote the unadjusted NPV as $\delta'(t_i)$ for transactions t_i. Then

$$\delta'(t_i) = \begin{cases} m \times v(t_i), & \text{if } \hat{y}(t_i) = 0, t_i \in S_{AG} \cup S_{RG}; \\ -(1-m) \times v(t_i), & \text{if } \hat{y}(t_i) = 0, t_i \in S_{AB} \cup S_{RB} \cup S_{RU}; \\ 0, & \text{if } \hat{y}(t_i) = 1. \end{cases} \tag{3.13}$$

3.6.2.1 Model Evaluation Criteria Performance

To evaluate the proposed model evaluation criterion, I first consider one iteration of Algorithm 2 for choosing tuning parameters based on validation data. The goal is to show that the performance of adjusted NPV under validation data have a similar pattern with NPV directly calculated under group B data.

First, I set the initial values $p_1 = 0.8, p_2 = 0.2$ and change weight z from 0 to 40 with step size 1. Figure 3.3 reports the adjusted NPV under validation data, the unadjusted NPV under validation data, and the NPV under group B data, respectively. A detailed report of NPV can be found in Appendix E. From Figure 3.3, we can see that the largest NPV under group B data is achieved at $z = 13$, which is close to $z = 14$ as the one getting the

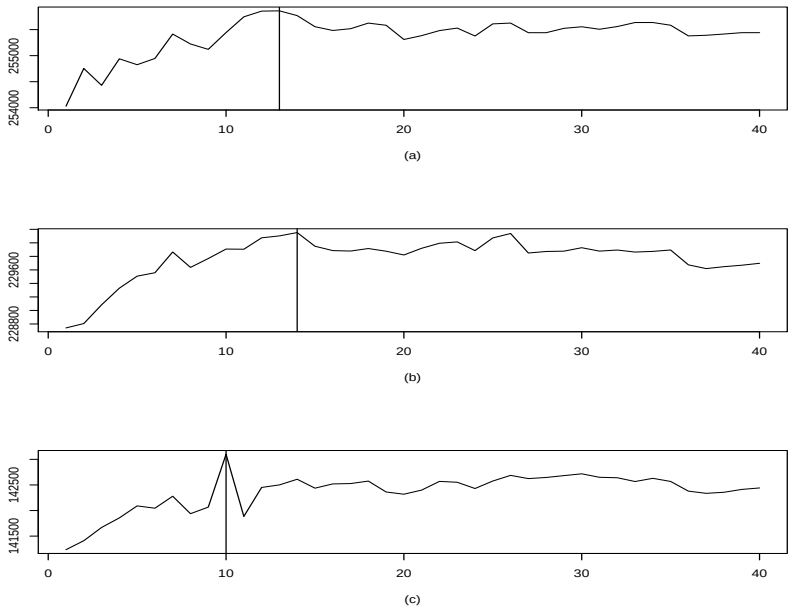


Figure 3.3: The NPV result for the proposed models with weight z varying from 0 to 40. (a) NPV under group B data; (b) Adjusted NPV under validation data; (c) Unadjusted NPV under validation data.

largest adjusted NPV under the validation data. In contrast, the largest unadjusted NPV under the validation data is achieved at $z = 10$. From Figure 3.3(c), the unadjusted NPV at $z = 10$ can be abnormal. However, if disregarding this point, the largest unadjusted NPV value is achieved at $z = 30$. Moreover, the trend of adjusted NPV under validation data is quite similar with the trend of NPV under group B data. But the trend of the unadjusted NPV under validation data does not share a similar pattern with the trend of NPV under group B data. In addition, the unadjusted NPV in (3.13) can be misleading. By treating all transactions in S_{RU} as bad transactions, the value of unadjusted NPV in Figure 3.3(c) has shrunk down significantly compared to the adjusted NPV in Figure 3.3(b). The largest value of unadjusted NPV is even lower than the online decision NPV value 210564.41 calculated using (3.8). Therefore, it is important to use the adjusted NPV in model evaluation.

Second, I set $z = 14$ for both group B and validation data and change the values of p_1 and p_2 from a grid where p_1 is set from 0 to 100% and p_2 is set from 0 to 70% both with a step size 5%. The results of NPV under group B data, adjusted NPV under validation data,

and unadjusted NPV under validation data are displayed in Figure 3.4 $\sim$ 3.6. From these three figures, we can see that the general trend of NPV calculated from group B data is very similar as the adjusted NPV calculated from the validation data. The largest NPV is obtained at $p_1 = 0.4, p_2 = 0.55$ under group B data, while the largest adjusted NPV is obtained at $p_1 = 0.5, p_2 = 0.55$ under the validation data. In contrast, the result for the unadjusted NPV under validation data is very different: the largest NPV is achieved at $p_1 = 0.3, p_2 = 0.6$ and the shape of NPV surface in Figure 3.6 is also different from those in Figure 3.4 and Figure 3.5.

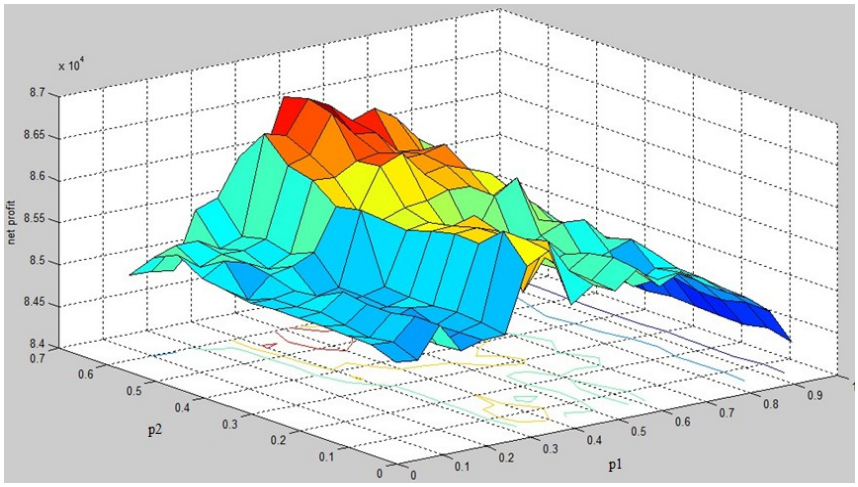


Figure 3.4: NPV result for models with combinations of p_1 and p_2 on Group B data

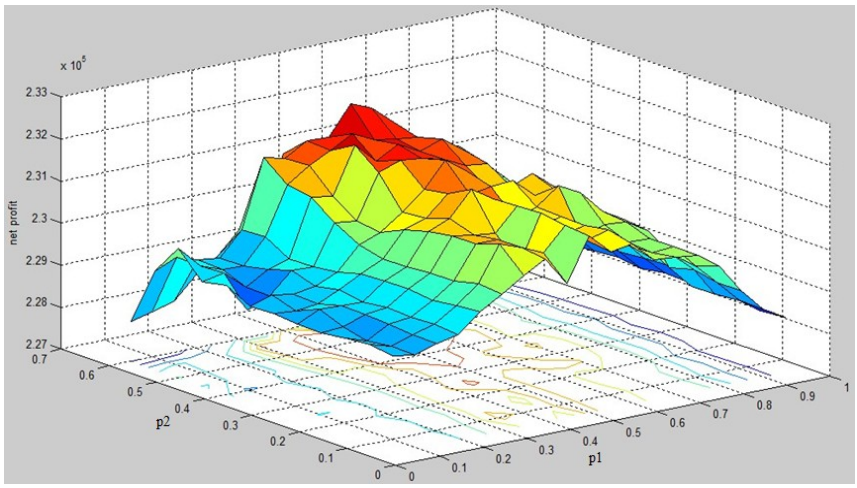


Figure 3.5: Adjusted NPV result for models with combinations of p_1 and p_2 on validation data

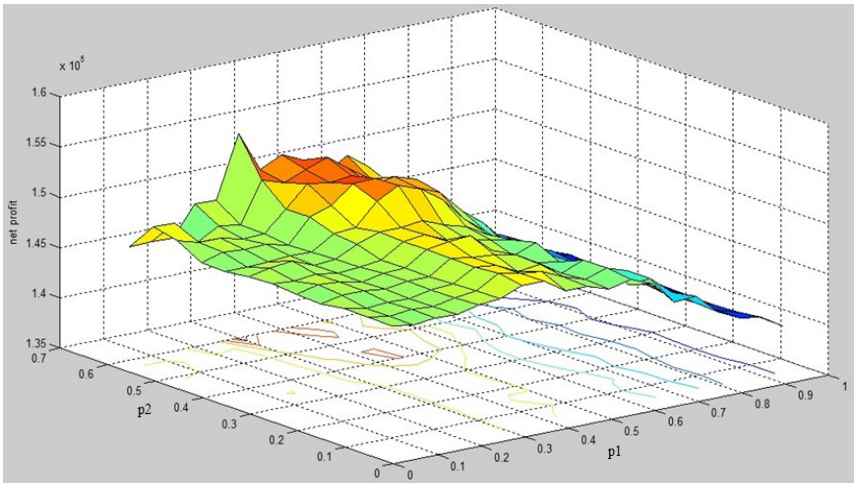


Figure 3.6: Unadjusted NPV result for models with combinations of p_1 and p_2 on validation data

Table 3.1: Model selection result with initial setting $p_1 = 0.8$ and $p_2 = 0.2$

Under Group B Data

Resultant Model	Tuning Parameters	NPV
Optimal model from iteration 1 *Step 1*	$z = 13, p_1 = 0.8, p_2 = 0.2$	85286.69
Optimal model from iteration 1 *Step 2*	$z = 14, p_1 = 0.4, p_2 = 0.55$	86852.16
Optimal model until algorithm converge	$z = 24, p_1 = 0.4, p_2 = 0.55$	87032.85
Online Decision		66479.43

Under Validation Data

Resultant Model	Tuning Parameters	Adjusted NPV
Optimal model from iteration 1 *Step 1*	$z = 14, p_1 = 0.8, p_2 = 0.2$	230151.82
Optimal model from iteration 1 *Step 2*	$z = 14, p_1 = 0.5, p_2 = 0.55$	232372.13
Optimal model until algorithm converge	$z = 24, p_1 = 0.55, p_2 = 0.55$	232613.37
Online Decision		210564.41

Furthermore, I continue Algorithm 2 until it converges. Table 3.1 displays the results for final models selected from *Step 1*, *Step 2* in the first iteration and the final optimal model at convergence. Under both the validation data and group B data, Algorithm 2 converges fast in three iterations. The result of online decision is listed as a baseline here. It is seen that the tuning parameters selected under group B data and validation data are quite close to each other in each step. It indicates that the adjusted NPV under validation data performs similar as the NPV under group B data.

From Table 3.1, the final model under validation data has increased the NPV by 11.05%

Table 3.2: Comparison of the two-stage model with other methods

	Under Group B Data			**Under Validation Data**		
Method	NPV	FP	FN	Adjusted NPV	Adjusted FP	Adjusted FN
Proposed	86753.36	1831	102	232613.37	3167	598
M_1	81594.64	2527	97	220809.14	5041	541
M_2	83950.82	2043	123	223283.55	4012	652
M_3	82929.78	2303	96	222909.86	4451	560

compared with that of the baseline online decision. It implies that the proposed model can significantly improve the current online decision. Under group B data, the NPV from the proposed method has been increased by 30.92%, suggesting that the A/B testing technique is very costly. As the proposed adjusted NPV criterion can mimic the performance of NPV calculated from group B data, it can be used as a proper model evaluation criterion, serving a good alternative for the costly A/B testing.

3.6.2.2 Model Prediction Performance

This section is to compare the proposed two-stage risk model with three existing models, including: M_1, the one-stage logistic regression with all rejected transactions removed; M_2, the one-stage logistic regression treating all the rejected transactions as bad transactions; M_3, the Heckman's two-stage model (1979) under the MNAR assumption. The adjusted NPV is used for the proposed model to choose the optimal tuning parameters under the validation data. To make a fair comparison, each model is run with the same rejection rate and they all go through the pre-screening step with a same threshold u. Under this circumstance, the proposed model has three more tuning parameters than the comparing models, z, τ_1 and τ_2.

The comparison results of NPV, FP and FN are reported in Table 3.2. From Table 3.2, it is clear that the proposed method outperforms the other methods in terms of NPV and FP. Although the number of FN in the proposed method is slightly larger compared with that in Model M_1 and Model M_3, the total number of false selection (FP+FN) is much smaller in the proposed method.

To further evaluate the prediction performance of the proposed method, I randomly sample

80% of the original group B data as a new test data for comparing the models. This sampling procedure is repeated 50 times. Figure 3.7 shows the boxplots for the values of NPV, FP and FN. The results further confirm that the NPV of my proposed two-stage model is much higher than that of the other three methods. From Figure 3.7(b) and Figure 3.7(c), we can see that the proposed method performs best as it largely reduces the FP numbers while its FN result is comparable with the other models. It is worth pointing out that Model M_1, removing all the rejected transactions, gives the worst performance in terms of the NPV values and FP numbers. It confirms the importance of using reject inference.

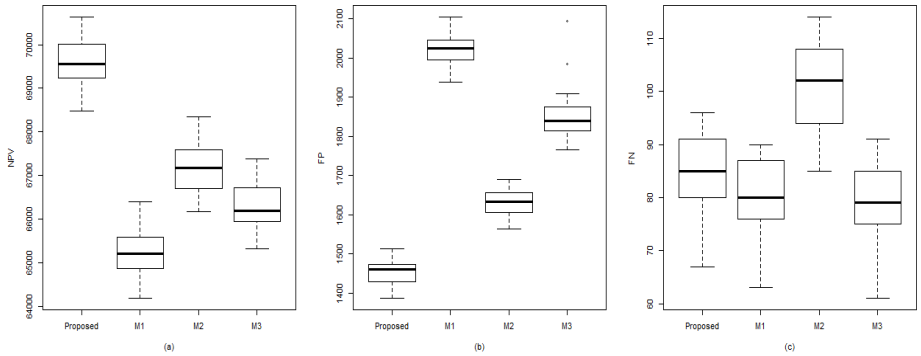


Figure 3.7: (a) NPV result under randomized group B data; (b) FP result under randomized group B data; (c) FN result under randomized group B data.

3.6.2.3 Result from Another Initial Setting

As seen from Algorithm 2, selection of tuning parameters depends on the initial values of p_1 and p_2. To examine the robustness of the proposed method with respect to the choice of initial settings, I consider using another initial setting $p_1 = 0.1, p_2 = 0.3$. Table 3.3 displays the optimal tuning parameters selected in the first iteration after each step of Algorithm 2 as well as the optimal model at convergence. The algorithm also converges fast in three iterations.

As stated in Section 3.4.2, the algorithm may converge to different local maximums with different initial settings. From Table 3.3 with initial setting $p_1 = 0.1, p_2 = 0.3$, the tuning parameters selected under group B data and validation data remain close to each other in

Table 3.3: Model selection result with initial setting $p_1 = 0.1$ and $p_2 = 0.3$

Under Group B Data

Resultant Model	Tuning Parameters	NPV
Optimal model from iteration 1 *Step 1*	$z = 7, p_1 = 0.1, p_2 = 0.3$	85291.61
Optimal model from iteration 1 *Step 2*	$z = 7, p_1 = 0.4, p_2 = 0.5$	86910.24
Optimal model until algorithm converge	$z = 9, p_1 = 0.4, p_2 = 0.55$	86956.69
Online Decision		66479.43

Under Validation Data

Resultant Model	Tuning Parameters	Adjusted NPV
Optimal model from iteration 1 *Step 1*	$z = 7, p_1 = 0.1, p_2 = 0.3$	229527.78
Optimal model from iteration 1 *Step 2*	$z = 7, p_1 = 0.5, p_2 = 0.55$	232335.32
Optimal model until algorithm converge	$z = 9, p_1 = 0.5, p_2 = 0.55$	232797.49
Online Decision		210564.41

each step. The final model is obtained with $z = 9, p_1 = 0.5, p_2 = 0.55$ under validation data and $z = 9, p_1 = 0.4, p_2 = 0.55$ under group B data. Moreover, under validation data, the NPV of 232797.49 from the optimal model in Table 3.3 is also close to the NPV of 232613.37 from the optimal model in Table 3.1 with the initial setting $p_1 = 0.8, p_2 = 0.2$. Similar patterns are also observed for the FP and FN.

3.6.2.4 Results under a Randomized Setting

To further evaluate the proposed method, I also randomly select 80% of the original group A data as the new group A data. Similarly, 70% of the new group A data is used as new training data and the remaining 30% as new validation data.

First, the original group B data is used for test data. Results for the optimal model of each step for Algorithm is displayed in Table 3.4 and the model comparison results with other methods is displayed in Table 3.5. As seen from Table 3.4 and Table 3.5, the results further confirm the findings similar with that under the original group A data. The tuning parameters selected under group B data and the new validation are close to each other in Table 3.4. The NPV calculated under optimal model in Table 3.4 for group B data, 87026.10, is close to the one from Table 3.1, 87032.85. In Table 3.5, the resultant NPV shows the proposed method outperforms the other three methods. The number of FP is also reduced largely. The number

Table 3.4: Model selection result under randomized group A data

Under Group B Data

Resultant Model	Tuning Parameters	NPV
Optimal model from iteration 1 *Step 1*	$z = 5, p_1 = 0.3, p_2 = 0.2$	85753.93
Optimal model from iteration 1 *Step 2*	$z = 5, p_1 = 0.4, p_2 = 0.55$	87026.10
Optimal model until algorithm converge	$z = 5, p_1 = 0.4, p_2 = 0.55$	87026.10
Online Decision		66479.43

Under Validation Data

Resultant Model	Tuning Parameters	Adjusted NPV
Optimal model from iteration 1 *Step 1*	$z = 5, p_1 = 0.3, p_2 = 0.2$	184452.80
Optimal model from iteration 1 *Step 2*	$z = 5, p_1 = 0.4, p_2 = 0.55$	186636.74
Optimal model until algorithm converge	$z = 5, p_1 = 0.4, p_2 = 0.55$	186636.74
Online Decision		168092.37

Table 3.5: Comparison of the two-stage model under randomized group A data

	Under Group B Data			**Under Validation Data**		
Method	NPV	FP	FN	Adjusted NPV	Adjusted FP	Adjusted FN
Proposed	87026.10	1817	104	186636.74	2463	510
M_1	81198.00	2563	97	176630.89	4142	442
M_2	83973.71	2050	122	179496.31	3252	526
M_3	82933.95	2296	94	178995.93	3584	455

of FN is comparable for the proposed method in comparison with the other three methods. Because of the constraint on computational time, I have not conducted multiple randomized setting on group A data.

Second, I evaluate the proposed method by using this new group A data but randomly sampling 80% of the original group B data as the new testing data. The sampling process is repeated 50 times and the NPV, FP, FN are compared respectively with the other three methods using boxplots. Detailed boxplots for NPV, FP and FN are provided in Appendix F. We can find that the proposed model still provides the largest NPV with lowest FP and comparable FN comparing with the other three methods.

3.7 Discussion

In this chapter, I proposed a two-stage risk modeling for reject inference. The proposed method enables to make the acceptance/rejection decisions in both stages. The first stage model aims at detecting the already captured fraud pattern and the missing fraud simultaneously and the second stage is designed to further capture the fraud not detected in the first stage. I also proposed an adjusted NPV as a novel model comparison criterion. By investigating the key insight under the reject inference framework, the adjusted NPV can accurately calculate the net profit for transactions with unknown status. The proposed method can potentially be an alternative for conducting the costly A/B testing.

For the two-stage model, the logistic regression is used as the major modeling technique for classification. Other classification approaches may be also applicable for the proposed two-stage model. It will be interesting to investigate how the weighting scheme can be adopted under other classification approaches. In the proposed stage-one model, weight z is kept same across all transactions that are missing fraud. Ways to incorporate adaptive weight for more efficient model building can be another interesting aspect for future research. For the greedy algorithm, it may converge to different local maximums under different initial settings. In practice, to avoid a local maximum, one can consider multiple initial settings. Then the optimal model can be selected among the resultant optimal models with best performance.

For the adjusted NPV criterion, one critical issue is to estimate the parameters r and s in (3.11). In this work, since the information of A/B testing on real data is avaiable, I take advantage of group B information to estimate r and s for calculating the adjusted NPV. When the A/B testing result is not available, the estimate of r and s in the adjusted NPV can also be obtained from business domain knowledge or the historical data. Moreover, we can develop an alternative to A/B testing to get the parameters r and s. The key idea is to replace the 100% accepted transactions in group B by a less expensive experiment. Specifically, I will first construct several models $M_1, M_2, \cdots, M_h$ under different settings of r and s. Then each model M_i will be used as online decision system for a small group of users. Finally by comparison the

estimated NPV and true NPV for each model, we can estimate r and s accurately. Research on this topic will reported elsewhere in the future.

Chapter 4 A Mixed Variance Component Model for Quantifying the Elastic Modulus of Nanomaterials

4.1 Introduction

One dimensional (1D) nanomaterials possess great mechanical properties with wide applications in many areas (Cui and Lieber, 2001; Stangl et al., 2004, Djurii et al., 2004). Compared with bulk materials, 1D nanomaterials shows advantages on many mechanical properties including the material tensile strength and elastic modulus (Yu et al., 2000). In order to advance the application of 1D nanomaterials, it is important to accurately quantify the mechanical properties of nanomatrials. This paper focuses on developing a statistical modelling method to accurate estimate the elastic modulus, one of the most important properties of nanomaterial, from the nano experiment data containing various noise and potential artifacts.

There are several challenges for quantifying the 1D nanomaterials. First, experimental measurement of the behaviour of the nanomaterial is difficult due to nanoscale manipulation and the need of high-accuracy devices. Moreover, the measurement error and systematic errors can be easily introduced during the measuring process and may be relatively large (Yu et al., 2000). A common technique for measuring the elastic modulus of 1D nanomaterials is based on the scanning electron microscope (SEM) (Bogner et al., 2006) and the atomic force microscope (AFM) (Yu et al., 2000). To perform such measurements, one can use an AFM tip to deform a 1D nanomaterial that is supported at two ends. The AFM tip pushes at certain locations along the 1D nanomaterial. Then the force displacement can be measured by images taken from the SEM and the AFM. By examining the force deformation at different locations on a nanomaterial, one can quantify the properties of a nanomaterial such as its elastic modulus based on the experimental data.

However, such experimental data often contain various noise, possible bias, and systematic

errors (Deng et al., 2009; Mai and Deng, 2010). It calls for new approaches to quantify the elastic deformation of the 1D nanomaterial. Mai and Wang (2006) apply the theory of the free-free beam model (Benham and Crawford, 1987) to obtain the estimate of elastic modulus for the Znic Oxide (ZnO) nanobelt experiment data. Their approach is based on the deformation of nanobelt under some fixed forces applied by the AFM tip. Three parameters for the AFM tip can affect the output of the experiment. To find out the best experiment conditions, Song et al. (2010) developed a nonlinear model to estimate the optimal parameter settings. There are also several statistical methods developed along the line of quantifying the properties of nanomaterials. In Wang et al. (2010), statistical sampling techniques are utilized to project SEM images to acquire some of the nanomaterial properties such as the length, diameter and density. A geometric model is proposed to estimate the nanomaterial length based on the projected images and a statistical imputation method is used to estimate the nanomaterial density. To control the synthesis process of nanomaterials, a multinomial generalized linear model is proposed to predict the nanomaterial morphologies status (Dasgupta et al., 2008; Dasgupta et al., 2011). Other statistical models to control the nanowire growth and the experimental design strategy for synthesising nanomaterials includes Zhu et al. (2014) and Xu and Huang (2014). To accurately quantify the nanomaterial properties, both statistical and physical modeling techniques are useful. In Deng et al. (2009), they proposed a novel statistics model combined with the FFBM theory to improve the estimate of elastic modulus for 1D nanobelts. Their approach considered multiple adjustment terms with a forward model selection technique to identify initial bias and systematical error that occurred in the experiment.

Motivated by Deng, et al. (2009), we propose a mixed variance component model to further improve the estimation accuracy of the elastic modulus for nanobelt experimental data. When conducting an experiment using scanning of the nanobelt by the AFM tip, the nanobelt is often scanned multiple times under a given force (Mai and Wang, 2006). The collected data thus contain multiple deformation curves given each force level. Instead of using the average deformation curves from all replications under each force level as in Deng

et al. (2009), we consider using all profiles data for estimating the elastic modulus. With multiple deformation curves under each force level in the data, there will be between group variance, within group variance and within profile variance. Thus, a mixed variance component model would be more appropriate for taking into account of the multiple variance structures. Several covariance structures are considered for modeling the within profile variance, including the auto-regressive and Gaussian covariance structure, which enhances the flexibility of the proposed method, leading to accurate estimation of the elastic modulus for the nanobelt. From the computational aspect, the algorithm proposed for estimation enjoys computational efficiency by decomposing the original variance matrix into smaller matrices in accordance with its block diagonal structure. Moreover, by accommodating potential bias as an adjustment factor into the proposed method, a group adaptive forward backward selection (GFoBa) is developed to select the significant adjustment factor. Thus, the systematic bias can be automatically filtered out for model estimation.

The remainder of this chapter is organized as follows. The real data is introduced in Section 4.2 and certain existing approaches are described there. Details of the proposed models with multiple error structures are described in Section 4.3. To perform variable selection in the proposed model, Section 4.4 introduces the group adaptive forward backward selection algorithm and related theories. In Section 4.5, we apply the proposed method under several simulation settings and compare with other existing methods. Real data problems are studied in Section 4.6. Section 4.7 includes discussion and extensions of this work.

4.2 Data and Existing Approaches

In this work, data were obtained from the bending profiles of the Zinc Oxide ZnO nanobelt (NB) described in Mai and Wang (2006), where a continuous scan of the ZnO NB is performed as illustrated in Figure 4.2(a) under different force levels. Specifically, a silicon substrate with long and parallel trend is prepared, upon which the NB can be placed. Then a certain force level is added using the AFM tip. The AFM tip is pushed into contact with the ZnO NB

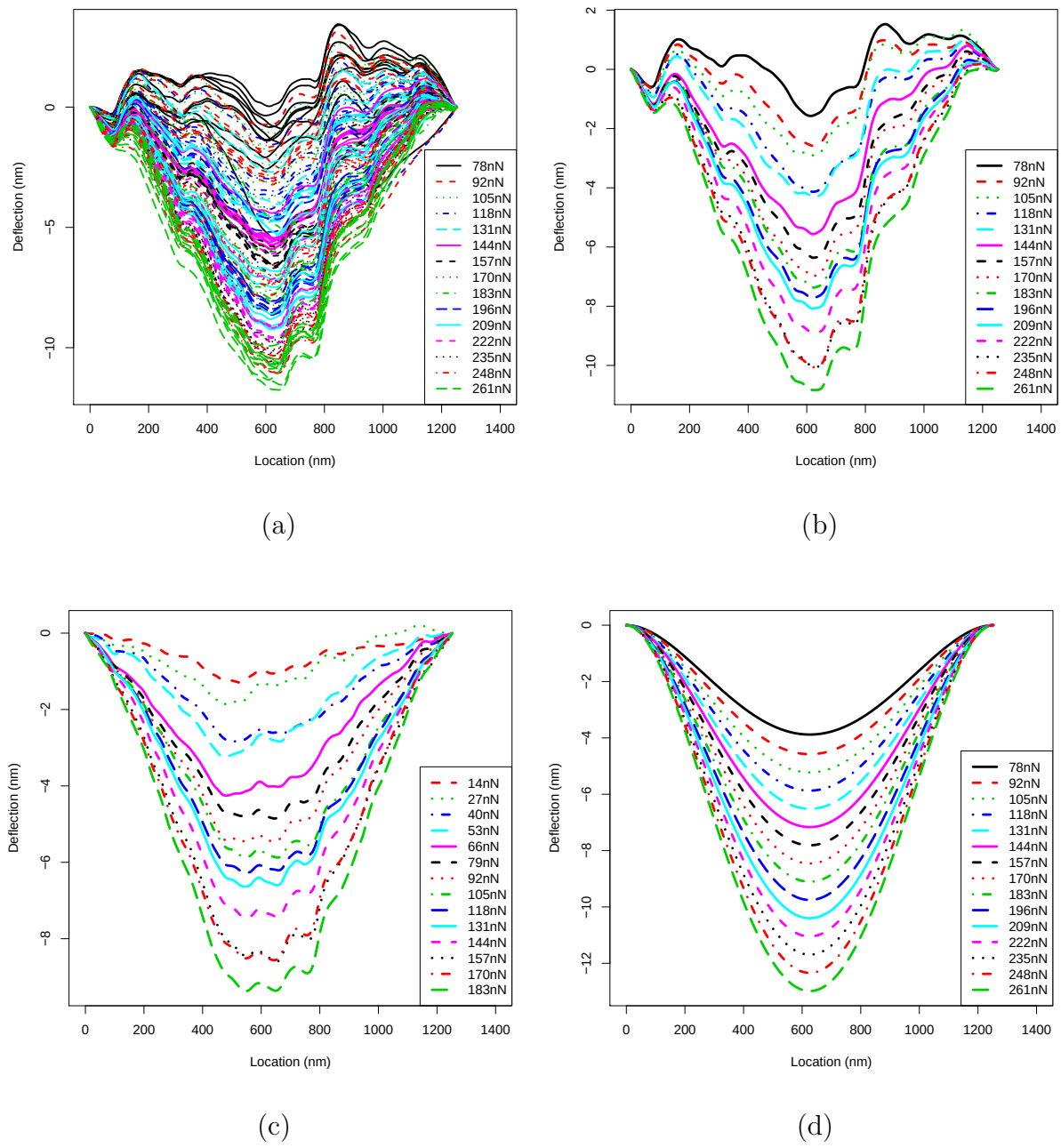


Figure 4.1: (a) Real data with 10 replicates under each of the 15 force levels. (b) The average bending profiles under each force level of the AFM tip. (c) The normalized average bending profiles by subtracting the average bending profile acquired at 78 nN from the profiles in (a). (d) The corresponding theoretical profiles of (a) under FFBM.

and moved along the length direction of NB. During the movement of the AFM tip, the SEM measures the displacement of the NB by taking images at multiple locations.

Based on the movement of the AFM tip from one end of NB to the other, the result of one

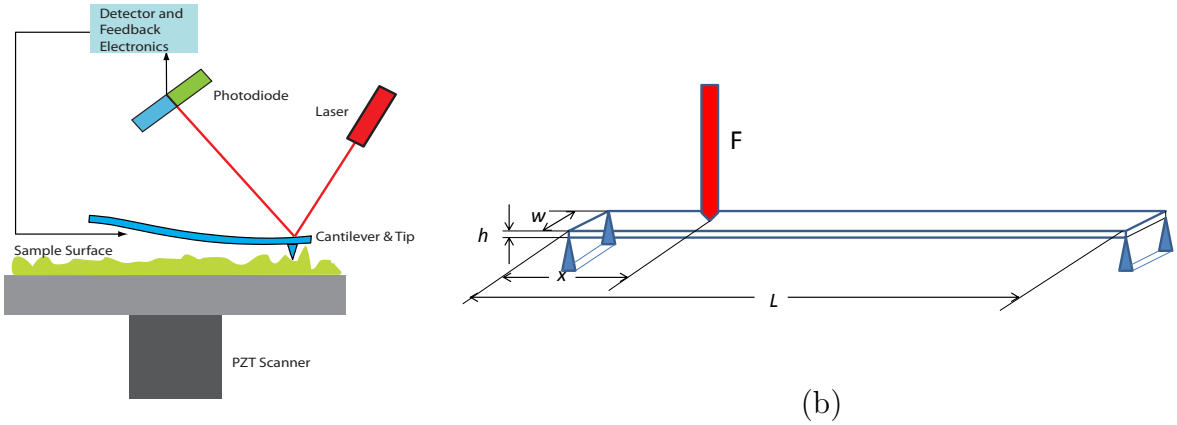
experiment is the NB bending profile data recording the measured displacements at multiple locations of the NB. The experiment is replicated at a given force level on the AFM tip for 10 times, producing 10 bending profiles of the NB at each force level. The magnitude of the force level is changed gradually from low to high at fixed levels, with 15 levels of force chosen from 78 nN (nano Newton) to 261 nN. As shown in Figure 4.1(a), there are thus totally 150 profiles recorded with 10 replications under each force level From 4.1(a), the deformation shapes are similar among 10 replications under a given force level. For the profiles recorded under different force levels, the profiles under large forces appears to have large deformation. Figure 4.1(b) displays the average profiles obtained by averaging the ten profiles under a same force level.

Since the adhesion between the ZnO NB and the silicon substrate is weak such that the two ends of the NB can slide freely, Mai and Wang (2006) considered to use the free-free beam model (FFBM) (Benham and Crawford, 1987) for estimating the elastic modulus based on data for averaged profiles in the experiment. The FFBM suggests that the displacement of the NB has a linear relationship with the force applied on the AFM tip:

$$v = -\frac{Fx^2(L-x)^2}{3EIL}, \tag{4.1}$$

where v is the displacement of NB at location x, E is the elastic modulus, L is the width of the trench, I is the moment of inertia given by $wh^3/12$ for the rectangular beam, where w and h are the width and thickness of NB, respectively. Figure 4.2(b) provides a illustrative schematic diagram of the FFBM.

The FFBM only considers the theoretical case where the experiments are performed without any noise factor and the smooth profiles can be obtained. Obviously, the curves obtained from real experiments are not smooth as illustrated in Figure 4.1(b). The irregular and non-smooth deformation curves can be due to the conditions of NB or the manipulation of AFM system. There is uneven surface roughness as well as initial bending in the NB. There are also multiple factors affecting the accuracy of the AFM system such as the size of the AFM tip and



(a) (b)

Figure 4.2: (a) The scanner system on the AFM tip. (b) Schematic diagram of the free-free beam model (FFBM).

the accuracy of positioning the AFM tip on the exact location. To eliminate the initial bias, Mai and Wang (2006) proposed to perform an adjustment by subtracting the initial average bending profile for all the subsequent profiles. As shown in Figure 4.1(c), such normalized deformation profiles appear to be smoother than the original profiles in Figure 4.1(b). Based on the normalized deformation profiles, the elastic modulus are then estimated based on the functional relationship between the elastic bending and the tip force applied according to (4.1).

It is noted that the normalizing step in Mai and Wang (2006) could potentially bring in more noise if the initial average bending profile is not accurate. Moreover, systematic biases could occur when changing the force level from one to the other (Deng et al., 2009). Using the data on averaged profiles as in Figure 4.1(b), Deng et al. (2009) proposed a sequential profile adjustment using a regression (SPAR) model extended from the FFBM scheme to alleviate initial bias and possible systematic errors. It thus can improve the estimation accuracy of the elastic modulus. The key idea in Deng et al. (2009) is to include initial bias as an adjustment factor in the model, and to accommodate systematic biases as multiple adjustment factors whenever the force changes. In such a way, their model can be more flexible than FFBM as the model accounts for noise and artifacts in the data. Since the system error may not always occur when the force level of AFM changes, a forward selection is performed to only choose

significant adjustment factors in the model.

In this work, we further generalize the method in Deng et al. (2009), enabling the analysis of all the profile replications data as shown in Figure 4.1(a). Specifically, we propose a mixed variance component model accounting for the variance structure generated by all profiles data, thus providing a more accurate estimate of the elastic modulus. Since there are 10 replications under each force level, the error structure can be further decomposed into several variance components, including between group variance, within group variance and within profile variance. Compared to Deng et al. (2009), the proposed method is more flexible and the variance decomposition can better characterize the variance structure of the data. We also propose a modified version of the adaptive forward backward selection (Zhang, 2009) to select significant adjustment terms in the model. Since all the profiles data are used, a group adaptive forward backward selection (GFoBa) is proposed such that an adjustment term selected will affect all the profile replications after the corresponding force change. Note that the forward selection procedure in Deng et al. (2009) could select insignificant terms since a selected term cannot be dropped even if it loses the predictive power after the subsequent terms are selected into the model (Kutner et al., 2005). In contrast of using GFoBa, the selected adjustment term will also have chance to be dropped if it is no longer necessary to remain in the model after subsequent adjustment terms are selected.

4.3 Proposed Model

The proposed method models all the profiles data as shown in Figure 4.1(a), which allows one to address the within group and between group variations as well as the correlation between locations within a profile itself.

Suppose the locations selected on a profile is fixed at $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ and the force level by $F_k, k = 1, \ldots, K$. Denote the observed displacement distance by v_{kij} at the jth location of the ith replicate under force level k. Let us denote by α_k, $k = 1, \cdots, K$, the between group variance for the kth force level applied. Denote by γ_{ki}, $i = 1, \cdots, M$, the within group

variance or the replication variance for the ith replication under each force level. Denote by ϵ_{kij}, $j = 1, \cdots, N$, the variance component for the jth fixed location on each profile. The proposed model includes all the three variance components to incorporate different sources of variances. Besides the multiple variance sources in the data obtained from the experiment, there may also exist initial bias and systematic bias introduced when changing from one force level to another. Such biases can influence all profiles thereafter once the bias occurs. Following a similar idea in Deng et al., (2009), adjustment terms $\delta_k, k = 1, \cdots, K-1$ are included in the model to adjust the systematic bias as well as δ_0 to adjust the initial bias and roughness condition of the NB. Thus, the proposed model for deflection at position x_j at replication i under the kth applied force level is as follows:

$$v_{kij} = f(x_j)F_k\beta_0 + \delta_0(x_j) + \delta_1(x_j)I(k > 1) + \cdots + \delta_{K-1}(x_j)I(k > K-1) + \alpha_k + \gamma_{ki} + \epsilon_{kij},$$

where $k = 1, \cdots, K, i = 1, \cdots, M, j = 1, \cdots, N$ and $I(\cdot)$ is an indicator function. Here the first term $f(x_j)F_k\beta_0$ is a re-expression of FFBM in (4.1) as a linear form of $\beta_0 = 1/E$ where E is the elastic modulus. The $f(x_j) = -\frac{x_j^2(L-x_j)^2}{3IL}$ where I, L are two fixed constants given by the experiment setting and the theoretical FFBM model. From the original FFBM model, the proposed model further incorporates various adjustment factors to calibrate possible systematic experimental errors through $\delta_0(x_j), \cdots, \delta_{K-1}(x_j)$ terms. Thus it combines theoretical and empirical models together to account for various noise and artifacts. Through a selection of the adjustment terms $\delta_1(x_j), \cdots, \delta_{K-1}(x_j)$, the proposed model makes minimal changes to the FFBM model (Chang and Joseph, 2014). Moreover, the proposed method flexibly integrates the engineering model and the statistical model by including the mixed variance components to account for randomness of the experiment (Joseph and Melkote, 2009; Joseph and Yan, 2014).

The indicate function $I(\cdot)$ is used to determine which of the adjustment terms should be included. Specifically, under the kth force level, there are in total $k-1$ times when the force level changes and all the corresponding adjustment terms need to be included such that the

model becomes:

$$v_{kij} = f(x_j)F_k\beta_0 + \delta_0(x_j) + \sum_{h=1}^{k-1}\delta_h(x_j) + \alpha_k + \gamma_{ki} + \epsilon_{kij}. \tag{4.2}$$

For the three variance components, we assume that they are independent with zero means and $\mathrm{Var}(\alpha_k) = \sigma_1^2$, $\mathrm{Var}(\gamma_{ki}) = \sigma_2^2$, $\mathrm{Var}(\epsilon_{kij}) = \tau_k^2$. Furthermore, we assume that there is no correlation between data points on different profiles. Under these assumptions, it is easy to see that $\mathrm{Corr}(v_{kij}, v_{k'i'j'}) = 0$ when $k \neq k'$. For the variance and covariance structure of ϵ_{kij}, denoted by $\boldsymbol{C}_k = (c_{ij}^{(k)})_{N\times N}$ under the force level F_k, we consider three cases Namely, there are (I) the independent setting, (II) the AR(1) setting: the autoregressive structure with order one, and (III) the Gaussian setting (Wolfinger, 1993; Liang and Zeger, 1986).

(I): The independent setting. In this setting, we assume that $\boldsymbol{C}_k = \tau_k^2\boldsymbol{I}$ with $\tau_k^2 = \sigma_3^2$. That is, the correlation of ϵ_{kij}'s for the data points on a same profile are zeros. Then it is easy to obtain that the correlation structure for the data v_{kij}'s as

$$\mathrm{Corr}(v_{kij}, v_{k'i'j'}) = \begin{cases} \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2}, & \text{if } k = k', i \neq i'. \\ \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2}, & \text{if } k = k', i = i', j \neq j'. \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

(II): The AR(1) setting. In this setting, we assume that $\boldsymbol{C}_k = \tau_k^2\boldsymbol{P}$. where $\tau_k^2 = \sigma_3^2$ and $\boldsymbol{P} = (p_{ij})_{N\times N}$ with $p_{ij} = \rho^{|x_i - x_j|}$. The $0 \leq \rho \leq 1$ is the lag 1 correlation under the AR(1) structure. It means that, for any two data points under the same profile, their correlation will be weaker if their positions are close with each other. Then the correlation structure for the

data v_{kij}'s can be written as

$$\text{Corr}(v_{kij}, v_{k'i'j'}) = \begin{cases} \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2+\sigma_3^2}, & \text{if } k = k', i \neq i'. \\ \frac{\sigma_1^2+\sigma_2^2+\sigma_3^2\rho^{|j-j'|}}{\sigma_1^2+\sigma_2^2+\sigma_3^2}, & \text{if } k = k', i = i', j \neq j'. \\ 0 & \text{otherwise.} \end{cases}$$

(III): The Gaussian setting. Note that $\boldsymbol{C}_1 = \cdots = C_K$ in the settings of I and II. In the setting of (III), we assume that $\boldsymbol{C}_k = \tau_k^2 \boldsymbol{R}$ where $\boldsymbol{R} = (r_{ij})_{N\times N}$ with $r_{ij} = \exp(-\lambda(x_i - x_j)^2)$. It means that the overall variance τ_k^2 is different for each force level. Here $0 \leq \lambda \leq 1$ is the correlation parameter. Then the correlation structure for the data v_{kij}'s can be written as

$$\text{Corr}(v_{kij}, v_{k'i'j'}) = \begin{cases} \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2+\tau_k^2}, & \text{if } k = k', i \neq i'. \\ \frac{\sigma_1^2+\sigma_2^2+\tau_k^2 exp(-\lambda(x_j-x_{j'})^2)}{\sigma_1^2+\sigma_2^2+\tau_k^2}, & \text{if } k = k', i = i', j \neq j'. \\ 0 & \text{otherwise.} \end{cases}$$

Note that the correlation between data points under the Gaussian setting can be easily adjusted to any one dimensional spatial structure.

Let us denote $\boldsymbol{v}_{ki}(\boldsymbol{x}, F_k) = (v_{ki1}, \ldots, v_{kiN})'$ as a vector of the deformation for all the data points of N locations on each profile. By stacking all the profiles data points, the proposed

model (4.2) can be transformed into the matrix form as:

$$\begin{pmatrix} \\ \boldsymbol{v}_{11}(\boldsymbol{x}, F_1) \\ \\ \vdots \\ \\ \boldsymbol{v}_{1M}(\boldsymbol{x}, F_1) \\ \boldsymbol{v}_{21}(\boldsymbol{x}, F_2) \\ \\ \vdots \\ \\ \boldsymbol{v}_{KM}(\boldsymbol{x}, F_K) \\ \\ \end{pmatrix} = \begin{pmatrix} f(\boldsymbol{x})F_1 & \boldsymbol{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f(\boldsymbol{x})F_1 & \boldsymbol{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ f(\boldsymbol{x})F_2 & \boldsymbol{I} & \boldsymbol{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f(\boldsymbol{x})F_K & \boldsymbol{I} & \boldsymbol{I} & \boldsymbol{I} & \cdots & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \\ \beta_0 \\ \\ \boldsymbol{\delta}_0 \\ \\ \vdots \\ \\ \boldsymbol{\delta}_{K-1} \\ \\ \end{pmatrix} + \boldsymbol{\xi}, \tag{4.4}$$

where $\boldsymbol{\delta}_k = \{\delta_k(x_1), \cdots, \delta_k(x_N)\}^T, k = 0, \cdots, K-1$. Here $\boldsymbol{\xi}$ is the corresponding error vector following $N(\mathbf{0}, \boldsymbol{\Sigma})$. The structure of $\boldsymbol{\Sigma}$, which is a $KMN \times KMN$ block diagonal matrix, can be obtained under the above three setting respectively. Denote by $\boldsymbol{B}_k$ which is a $MN \times MN$ matrix for the variance covariance matrix of the data points in a group of M profiles under a given force level F_k. Then $\boldsymbol{\Sigma}$ can be written as $\boldsymbol{\Sigma} = diag(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_K)$, where $\boldsymbol{B}_k$ can be expressed in general as

$$\boldsymbol{B}_k = \mathrm{Cov}(v_{k\cdot\cdot}) = \sigma_1^2 \boldsymbol{J}_{MN\times MN} + \sigma_2^2 \boldsymbol{I}_{M\times M} \otimes \boldsymbol{J}_{N\times N} + \boldsymbol{I}_{M\times M} \otimes \boldsymbol{C}_k,$$

where $\otimes$ is kronecker product and $\boldsymbol{J}$ is matrix with all elements equal to 1. Noting that under setting (I) and (II), $\boldsymbol{B}_1 = \cdots = \boldsymbol{B}_K$ since the within profile correlation structure is the same across all groups.

4.4 Estimation and Variable Selection

Generally, the proposed mixed variance components model in (4.4) can be written as a regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\xi}$, where $\boldsymbol{y}$ is the response vector, $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}$ is

the $(1 + K \times N)$ parameters vector $\boldsymbol{\beta} = \{\beta_0, \boldsymbol{\delta}_0, \cdots, \boldsymbol{\delta}_{k-1}\}^T$, and $\boldsymbol{\xi} \sim N(0, \boldsymbol{\Sigma})$. Denote by $\boldsymbol{\theta}$ the parameters in $\boldsymbol{\Sigma}$. Through the maximum likelihood approach, we propose an efficient iterative algorithm to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

Note that $\boldsymbol{\Sigma}$ is block diagonal matrix with $\boldsymbol{B}_k, k = 1, \cdots, K$ on its diagonal. Therefore, the log-likelihood function based on the proposed model in (4.4) can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) \propto -\sum_{k=1}^{K} [\log |\boldsymbol{B}_k| + (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})' \boldsymbol{B}_k^{-1} (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})], \tag{4.5}$$

where $\boldsymbol{y}_k$ is the deformation vector and $\boldsymbol{X}_k$ is the $MN \times MN$ design matrix for M profile replications under the kth force level. Given $\boldsymbol{\theta}$, we can have an explicit solution of $\boldsymbol{\beta}$ to maximize (4.5) as $\hat{\boldsymbol{\beta}} = (\sum_{k=1}^{K} \boldsymbol{X}_k' \hat{\boldsymbol{B}}_k^{-1} \boldsymbol{X}_k)^{-1} (\sum_{k=1}^{K} \boldsymbol{X}_k' \hat{\boldsymbol{B}}_k^{-1} \boldsymbol{y}_k)$ Given $\boldsymbol{\beta}$, the optimization of maximize (4.5) with respect to $\boldsymbol{\theta}$ is not trivial. Note that $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \sigma_3^2)' \triangleq \boldsymbol{\theta}_e$ in the setting (I) and $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \rho)' \triangleq \boldsymbol{\theta}_e$ in the setting (II). In the setting (III), $\boldsymbol{\theta} = (\sigma_1^2, \sigma_2^2, \lambda, \tau_1, \ldots, \tau_K)' \triangleq (\boldsymbol{\theta}_e, \tau_1, \ldots, \tau_K)$. There can be a large number of parameters for $\boldsymbol{\theta}$ in $\boldsymbol{B}_k$ in the setting of (III) for the Gaussian correlation structure. A direct optimization over $\boldsymbol{\theta}$ can be computationally expensive. To address this challenge, one can further estimate the parameters in $\boldsymbol{\theta}$ in an iterative fashion. The algorithm of estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is described as Algorithm 3.

Algorithm 3

Step 0: *Set the initial value of* $\boldsymbol{\theta}$ *as* $\boldsymbol{\theta}_0$.

Step 1: *Given* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, *update the value of* $\hat{\boldsymbol{\beta}}$ *as* $(\sum_{k=1}^{K} \boldsymbol{X}_k' \hat{\boldsymbol{B}}_k^{-1} \boldsymbol{X}_k)^{-1} (\sum_{k=1}^{K} \boldsymbol{X}_k' \hat{\boldsymbol{B}}_k^{-1} \boldsymbol{y}_k)$.

Step 2: *Given* $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, *solve for* $\hat{\boldsymbol{\theta}}_e$ *by* $\hat{\boldsymbol{\theta}}_e = \arg\max_{\hat{\boldsymbol{\theta}}_e} l(\boldsymbol{\beta}, \boldsymbol{\theta})$.

Step 3: *Given* $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\theta}_e = \hat{\boldsymbol{\theta}}_e$, *solve for* $\hat{\tau}_k^2 = \arg\max_{\tau_k^2} \log |\boldsymbol{B}_k| + (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})' \boldsymbol{B}_k^{-1} (\boldsymbol{y}_k - \boldsymbol{X}_k\boldsymbol{\beta})$, $k = 1, \cdots, K$.

Step 4: *Go back to* Step 1 *and repeat* Step 1 *to* Step 3 *until all parameters converge.*

The stopping criteria for Algorithm 3 are $||\hat{\boldsymbol{\theta}}_{e,t} - \hat{\boldsymbol{\theta}}_{e,t-1}||_2^2 < \epsilon_1$ and $||\hat{\boldsymbol{\beta}}_t - \hat{\boldsymbol{\beta}}_{t-1}||_2^2 < \epsilon_2$ where $\hat{\boldsymbol{\theta}}_{e,t}$, $\hat{\boldsymbol{\beta}}_t$, $\hat{\boldsymbol{\theta}}_{e,t-1}$ and $\hat{\boldsymbol{\beta}}_{t-1}$ are the estimates of $\boldsymbol{\theta}_e$ and $\boldsymbol{\beta}$ in the tth and $(t-1)$th iteration, ϵ_1

and ϵ_2 are two pre-selected small values. For the settings of (I) and (II), Step 3 of Algorithm 3 can be omitted.

Note that the adjustment factors $\boldsymbol{\delta}_k$ for systematic bias may not always be significant. That is, only a few $\boldsymbol{\delta}_k$ terms are needed in model (4.4). To effectively estimate the parameters and enhance the accuracy of the model, we conduct a variable selection procedure for $\boldsymbol{\delta}_k, k = 1, \ldots K-1$. Since each $\boldsymbol{\delta}_k$ is a N-dimensional vector, a group variable selection would be more appropriate and should evaluate which parameters in a $\boldsymbol{\delta}_k$ vector need to be included or excluded at the same time. Motivated by the adaptive forward backward selection in Zhang (2009), we develop a group adaptive forward backward selection (GFoBa) procedure to choose significant $\boldsymbol{\delta}_k$'s. The GFoBa procedure is described as follows, where f_s is defined as the group variable selection criterion function which will be discussed later.

GFoBa Procedure (for Variable Selection)

Step 1. Select one group of parameters $\boldsymbol{\delta}_{i^*}$ where $i^* = \arg\min_i f_s(\boldsymbol{\delta}_i)$. Denote the active parameters set as $S^{(h_1)} = \{\boldsymbol{\delta}_{i^*}\}$.

Step 2. A forward selection step: select $\boldsymbol{\delta}_{j^*}$ from the remaining parameters groups where $j^* = \arg\min_j f_s(S^{(h_1)} \cup \boldsymbol{\delta}_j)$. Update the active parameters set as $S^{(h_2)} = \{S^{(h_1)} \cup \boldsymbol{\delta}_{j^*}\}$.

Step 3. A backward selection step: find $\boldsymbol{\delta}_{k^*}$ where $k^* = \arg\min_k f_s(S^{(h_2)} \setminus \boldsymbol{\delta}_k)$. If $[f_s(S^{(h_2)}) - f_s(S^{(h_2)} \setminus \boldsymbol{\delta}_k)] < \alpha[f_s(S^{(h_1)}) - f_s(S^{(h_2)})]$, drop $\boldsymbol{\delta}_{k^*}$ from the active set $S^{(h_2)}$. Update the new active parameters set as $S^{(h_1)} = S^{(h_2)}$.

Step 4. Repeat *Step 2* and *Step 3* until no parameters group to be added or the change of f_s value less than a pre-specified ϵ_c.

Note that the value α in *Step 3* controls how likely it is that a parameter group is removed in a backward step. Here we set $\alpha = 0.5$ throughout the paper to moderately control the backward step. For the choice of variable selection criterion function f_s, we consider Akaike information criterion (AIC) and Bayesian information criterion (BIC) as two commonly used

criteria:

$$\text{AIC} = \log|\hat{\boldsymbol{\Sigma}}| + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + 2\text{df},$$
$$\text{BIC} = \log|\hat{\boldsymbol{\Sigma}}| + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + \text{df}\log n,$$

where $n = KMN$ is the total number of observations, df$= 1 + g \times N + ||\boldsymbol{\theta}||_0$ is the degrees of freedom of the corresponding model, where g is the number of significant groups selected from $\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_{K-1}$, $||\cdot||_0$ denotes the l_0 norm.

The proposed GFoBa procedure can provide accurate estimation for $\boldsymbol{\beta}$. We investigate the consistency property of GFoBa for the proposed mixed variance component model. Following Zhang (2009), several assumptions are needed as follows:

(A1) Not all the $\boldsymbol{\delta}_k$'s are significant in the variable set $\boldsymbol{\beta}$ such that there exist a subset $\bar{\boldsymbol{\beta}}$ in $\boldsymbol{\beta}$ that satisfies the assumption $\boldsymbol{X}\bar{\boldsymbol{\beta}} = E(\boldsymbol{y})$;

(A2) The response $y_i, i = 1, \cdots, n$ are independent sub-Gaussians: there exists $\sigma^2 \geq 0$ such that $\forall i, t \in R$, $E\left(e^{t(y_i - Ey_i)}\right) \leq e^{\sigma^2 t^2/2}$.

(A3) Consider $\bar{\boldsymbol{\beta}} \in R^d$ and assume that in the GFoBa procedure $\epsilon_c \geq 108\rho(s)^{-1}\sigma^2 \ln(16d)/n$, where $s \leq d$ such that $32(\bar{k} + 1) \leq (0.8s - 2\bar{k})\rho(s)^2$ with $\bar{k} = |\text{supp}(\bar{\boldsymbol{\beta}})|$ and $\rho(k) = \inf\{\frac{1}{n}||\boldsymbol{X}\boldsymbol{\beta}||_2^2/||\boldsymbol{\beta}||_2^2 : ||\boldsymbol{\beta}||_0 \leq k\}$.

Theorem 1 *Under assumptions (A1)-(A3), for the mixed variance component model, the GFoBa can ensure the convergence of $\hat{\boldsymbol{\beta}}$ given the maximum likelihood estimate $\hat{\boldsymbol{\Sigma}}$. That is, when GFoBa procedure ends at step $k \leq s - \bar{k}$, the estimated $\hat{\boldsymbol{\beta}}$ denoted as $\boldsymbol{\beta}^{(k)}$ satisfies:*

$$(i). \quad ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}E\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}E\boldsymbol{y}||_2^2 \leq g_1(k, n);$$
$$(ii). \quad ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}E\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}E\boldsymbol{y}||_2^2 \leq g_2(k, n),$$

where $\bar{\boldsymbol{\beta}}$ is the true mean vector, n is the total number of observations, $g_1(k, n)$, $g_2(k, n)$ are functions of k and n and they go to 0 as $n \to \infty$.

From Theorem 1 we can see that the estimate of $\boldsymbol{\beta}$ from GFoBa enjoys consistency properties.

Given the maximum likelihood estimate of $\hat{\boldsymbol{\Sigma}}$, the estimated $\boldsymbol{\beta}^{(k)}$ stopping at the kth step of the GFoBa procedure will be close to the true $\bar{\boldsymbol{\beta}}$ as n increases. From Theorem 1(ii), it can be further proved that the difference between $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)}$ and $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}$ can even be bounded above by the corresponding difference given the true $\boldsymbol{\Sigma}$ and $\bar{\boldsymbol{\beta}}$ as n increases. The detailed proof of Theorem 1 is provided in Appendix G.

4.5 Simulations

To examine the performance of the proposed method, we conduct a simulation study by using data generated according to the FFBM model (4.1). Following the real data case in Figure 4.1(a), we consider $K = 15$ groups of deformation curves with $M = 10$ replications in each group at force levels $78, 92, \ldots, 248, 261$ nN, respectively.

To generate the simulation data, we follow the proposed model in (4.4). The deformation curve are recorded at $N = 30$ equally spaced locations of the NB. The constants in the FFBM are assumed to be the same as in the real experiment in Mai and Wang (2006). The true elastic modulus value is set at $E = 100$. For the covariance structure $\boldsymbol{\Sigma}$ of the error vector $\boldsymbol{\xi}$, we consider the three settings as the described in Section 4.3. For the setting (I), the parameters are set as $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.2, 0.1, 0.3)$. For the setting (II), the values of $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ are set the same as in setting (I) and the correlation parameter $\rho = 0.5$. For the setting (III), the values of (σ_1^2, σ_2^2) are set the same as in setting (I) and the correlation parameter $\lambda = 0.7$. The within profile variance $\tau_1^2, \cdots, \tau_{15}^2$ are randomly generated from a uniform distribution U(0.15,0.35).

To mimic the real experiment that some systematic biases may occur in the experiment, we consider only a few of $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{K-1}$ to be significant with non-zero coefficients. Those significant $\boldsymbol{\delta}_k$'s could cause the shifting over the theoretical deformation curves. Specifically, three schemes are considered by randomly choosing: (1) none of $\boldsymbol{\delta}_k$'s being significant, (2) two of $\boldsymbol{\delta}_k$'s being significant, and (3) five of $\boldsymbol{\delta}_k$'s being significant. The significant $\boldsymbol{\delta}_k$'s also follow the pattern of the FFBM model (4.1). The size of the $\boldsymbol{\delta}_k$'s are carefully controlled by imposing a small random force upon (4.1) with the force level generated from a uniform

distribution U(3,13). The range is chosen to ensure that the significant $\boldsymbol{\delta}_k$ groups are shifted significantly but not shifted too much. Then a random noise vector generated from $N(\mathbf{0}, \sigma_0^2 \boldsymbol{I})$, where $\sigma_0 = 0.1$, is added on each of the significant $\boldsymbol{\delta}_k$ profile as the final $\boldsymbol{\delta}_k$.

Under each simulation setting, we generate the data 50 times and perform the proposed method for model estimation and variable selection. Besides the proposed mixed variance component model, the SPAR method in Deng et al. (2009) and the MW method in Mai and Wang (2006) are also performed for comparison purpose. Several performance measures are used to compare the three methods. We use the root mean square error (RMSE) (Hyndman and Koehler, 2006) to examine the fitting performance of the model:

$$\text{RMSE} = \sqrt{\sum_k \sum_i \sum_j (v_{kij} - \hat{v}_{kij})^2 / n}, \tag{4.6}$$

where $n = KMN$ is the total number of observations. For the accuracy of the group variable selection, the averaged true positive (TP), false positive (FP) are reported to evaluate whether the selected groups are the same with the true significant groups. Note that the true negative (TN) and false negative (FN) values are not reported for conciseness since the summation of TP and FN adds up to the total number of significant $\boldsymbol{\delta}_k$ groups corresponding to the three settings of $\boldsymbol{\delta}_k$'s, and the summation of FP and TN adds up to the total number of groups which is $K = 15$. A key objective is to accurately estimate the elastic modulus E of the NB. Based on the delta method (Casella and Berger, 2002), we can estimate the elastic modulus accordingly since $\hat{E} = 1/\hat{\beta}_0$. A good estimate of E should have small bias as well as small variance. Thus we calculate the mean square error $\text{MSE}(\hat{E}) = \text{Var}(\hat{E}) + \text{Bias}(\hat{E}, E)^2$ as a measure of accuracy of $\hat{E}$.

Tables 4.1 - 4.3 display the comparison results under different settings of $\boldsymbol{\Sigma}$, respectively. The proposed mixed variance component model under the three settings of (I), (II) and (III) are denoted as Proposed-I, Proposed-II and Proposed-III, respectively. Several performance measures including the RMSE, $\hat{E}$, $\text{MSE}(\hat{E})$, TP and FP are reported with their standard errors displayed in parentheses. As a baseline, the performance measures on the generated

original data are also reported including the RMSE, the true E, TP and FP. Specifically, the RMSE from the original data is calculated based on (4.6) with the $\hat{v}_{kij}$ obtained from the theoretical model (4.1) using the true E. It can be seen that the proposed methods outperform SPAR and MW method in terms of fitting accuracy according to the RMSE. Among which the proposed method with the Gaussian setting usually yields the lower RMSE however the data is generated. Under most cases, the RMSE from the proposed method are of the same size as the baseline RMSE. There are two cases for RMSE where it differs largely from the baseline RMSE under original data. For example, in Table 4.1, under the Gaussian setting assumption under AIC with no significant $\boldsymbol{\delta}_k$ case, the RMSE has a mean value of 0.53 which differs largely from the baseline RMSE 51.90. The large discrepancy is caused by the false positives that are introduced under the model with the assumption of setting (III).

Regarding the estimation accuracy of elastic modulus, the estimated $\hat{E}$ from the proposed method is closer with the true E compared with the estimated $\hat{E}$ from SPAR or MW method. For the proposed method under different settings of $\boldsymbol{\Sigma}$, generally the $\hat{E}$ will be more accurate when the model assumption matches the setting used for data generation. For example, the results in Table 4.3 show that the estimation under setting (III) is more accurate compared with that under setting (I) or (II). The reason is that the data is generated from the same setting which is exactly the assumption of setting (III). We also observe that the number of significant $\boldsymbol{\delta}_k$'s in the model can have large impact on the estimation accuracy. In general, as the number of significant $\boldsymbol{\delta}_k$'s increase, the estimation accuracy of the elastic modulus E decrease. Specifically, when the number of significant $\boldsymbol{\delta}_k$'s is zero or two, the estimated $\hat{E}$ is closer to the true value 100. While the estimated $\hat{E}$ is biased downward largely when there are five significant $\boldsymbol{\delta}_k$'s.

The variance of the estimated $\hat{E}$ from the proposed method are smaller compared with that of the SPAR and MW under most cases thus yielding a smaller MSE. The explanation is that the proposed methods use all replications information rather than just the averaging profiles such that the estimate of E is more stable and accurate. Note that some of the MSE($\hat{E}$) have large standard errors. For different simulation data sets, the selected $\boldsymbol{\delta}_k$'s may not be accurate

all the time such that the estimated $\hat{E}$ can be far away from the true E, which causes a large standard error of MSE($\hat{E}$).

For the group variable selection accuracy, the proposed GFoBa procedure works well under most cases with TP and FP, especially under setting (I) and (II). Compared with SPAR, the TP and FP of the proposed methods are closer to the true baseline TP value. The performance of TP differs when using AIC as the selection criterion compared with using BIC. When there are zero or two significant $\boldsymbol{\delta}_k$'s, all three settings (I), (II), (III) can identify the significant groups correctly most of the time for the proposed method. While Proposed-III tends to include more groups under AIC criterion when the simulation data is not generated from setting (III). For the comparison of using AIC or BIC for the proposed method, it is seen that AIC outperforms BIC regarding the estimation accuracy of E. Using AIC as the criterion also yields more stable estimates with smaller standard error. Although the AIC yields more FP, it also catch more TP among the significant $\boldsymbol{\delta}_k$'s. The BIC criterion chooses simpler model with less independent variables than AIC such that there are more groups included under AIC. When there are five significant $\boldsymbol{\delta}_k$'s, it is more difficult to select the right groups that are shifted. Models constructed under BIC usually selects only two or three groups in the final model while models under AIC usually will include more than five groups.

4.6 Real Data Analysis

As described in Section 4.2, the real data comes from the ZnO NB experiment containing 150 profiles in total with 10 replications under each of the 15 force levels. On each profile, there are 161 equal-spaced positions with the deformation data points recorded. We applied the proposed method for this real data. In comparison, the MW method (Mai and Wang, 2006) and the SPAR method (Deng et al., 2009) are also applied on the average profiles data.

Table 4.4 reports RMSE, $1/\hat{\mathrm{E}}$ and $\hat{E}$ with their standard errors under all compared methods. As described in Section 4.5, it appears that AIC works better than BIC on the estimation accuracy of the elastic modulus. Here the proposed method thus uses AIC as the selection

Table 4.1: Result for Data Generated from Setting (I)

# of Significant $\boldsymbol{\delta}_k$'s	Criterion	Model	RMSE	$\hat{E}$	MSE($\hat{E}$)	TP	FP
0	AIC	Proposed-I	51.29(0.48)	100.02(0.03)	0.12(0.01)	0.00(0.00)	0.00(0.00)
		Proposed-II	53.04(0.50)	100.02(0.04)	0.14(0.02)	0.00(0.00)	0.00(0.00)
		Proposed-III	0.53(0.01)	100.29(0.36)	15.67(3.38)	0.00(0.00)	2.92(0.04)
		SPAR	71.67(0.82)	100.10(0.11)	1.39(0.13)	0.00(0.00)	0.00(0.00)
	BIC	Proposed-I	51.29(0.48)	100.02(0.03)	0.12(0.01)	0.00(0.00)	0.00(0.00)
		Proposed-II	53.04(0.50)	100.02(0.04)	0.14(0.02)	0.00(0.00)	0.00(0.00)
		Proposed-III	25.98(3.59)	100.03(0.12)	3.75(0.92)	0.00(0.00)	0.96(0.16)
		SPAR	71.67(0.82)	100.10(0.11)	1.39(0.13)	0.00(0.00)	0.00(0.00)
	MW		10.73(0.28)	98.55(1.20)	72.68(1.61)	-	-
	Baseline		51.90(0.43)	100	-	0	0
2	AIC	Proposed-I	0.70(0.05)	98.15(0.85)	40.07(27.72)	1.78(0.08)	0.94(0.19)
		Proposed-II	0.70(0.05)	99.10(0.45)	12.15(7.67)	1.74(0.07)	1.04(0.19)
		Proposed-III	0.24(0.00)	97.21(0.91)	84.48(15.41)	1.72(0.06)	4.26(0.07)
		SPAR	6.68(0.47)	94.54(1.16)	97.75(29.17)	1.08(0.09)	0.52(0.12)
	BIC	Proposed-I	2.36(1.17)	96.60(0.66)	33.45(9.65)	1.34(0.09)	0.28(0.08)
		Proposed-II	1.18(0.07)	95.82(0.74)	44.96(12.02)	1.28(0.10)	0.36(0.07)
		Proposed-III	0.75(0.07)	96.04(0.77)	57.27(12.16)	1.48(0.08)	1.54(0.23)
		SPAR	9.22(0.26)	91.61(1.11)	33.45(9.65)	0.72(0.06)	0.52(0.12)
	MW		17.96(1.72)	73.07(1.60)	850.59(80.87)	-	-
	Baseline		0.85(0.01)	100	-	2	0
5	AIC	Proposed-I	0.25(0.02)	89.75(1.87)	284.81(66.81)	4.30(0.11)	2.52(0.28)
		Proposed-II	0.23(0.01)	91.06(2.00)	287.32(73.39)	4.34(0.11)	2.78(0.28)
		Proposed-III	0.14(0.00)	92.50(2.40)	619.96(90.36)	4.48(0.11)	5.44(0.12)
		SPAR	0.46(0.07)	83.60(2.84)	1103.11(107.07)	4.00(0.13)	5.61(0.21)
	BIC	Proposed-I	0.74(0.06)	79.68(1.67)	551.70(70.74)	2.24(0.17)	0.62(0.08)
		Proposed-II	0.93(0.07)	77.36(1.90)	689.96(94.74)	1.94(0.16)	0.50(0.09)
		Proposed-III	0.38(0.03)	80.44(2.27)	701.54(94.84)	3.06(0.18)	2.28(0.26)
		SPAR	1.28(0.05)	67.94(1.52)	1141.49(107.96)	0.64(0.07)	0.39(0.21)
	MW		23.45(0.80)	50.57(0.97)	2490.34(91.23)	-	-
	Baseline		0.34(0.00)	100	-	5	0

criterion. From the results in Table 4.4, it is seen that the proposed method under setting (I) yields the smallest standard error for $\hat{E}$. While the MW method yields the largest standard error for the estimated $\hat{E}$. For the RMSE, the proposed method usually yields a smaller RMSE compared with that of the SPAR or MW method using only the average profiles data. It implies that the mixed variance component can better capture the data variation. Note that the true elastic modulus is unknown in this real experiment. Based on the estimated $\hat{E}$ and its standard errors, the proposed method under the independent setting can give a more reliable estimate of elastic modulus than other approaches. It is worth pointing out that the proposed method under the AR(1) and Gaussian settings also have small RMSE with comparable standard errors with SPAR.

Table 4.2: Result for Data Generated from Setting (II)

# of Significant $\boldsymbol{\delta}_k$'s	Criterion	Model	RMSE	$\hat{\text{E}}$	MSE($\hat{\text{E}}$)	TP	FP
0	AIC	Proposed-I	49.71(1.49)	100.05(0.05)	0.21(0.03)	0.00(0.00)	0.04(0.03)
		Proposed-II	51.70(0.42)	99.98(0.05)	0.31(0.03)	0.00(0.00)	0.00(0.00)
		Proposed-III	3.62(1.64)	99.96(0.29)	10.48(1.71)	0.00(0.00)	2.34(0.13)
		SPAR	66.60(1.28)	99.99(0.20)	2.77(0.38)	0.0(0.00)	0.04(0.03)
	BIC	Proposed-I	51.64(0.49)	100.04(0.06)	0.21(0.03)	0.00(0.00)	0.00(0.00)
		Proposed-II	51.70(0.42)	99.98(0.05)	0.31(0.03)	0.00(0.00)	0.00(0.00)
		Proposed-III	51.63(0.52)	100.04(0.06)	0.7(0.03)	0.00(0.00)	0.00(0.00)
		SPAR	77.90(1.11)	99.98(0.19)	2.57(0.34)	0.00(0.00)	0.00(0.00)
	MW		11.05(0.29)	100.79(1.16)	6.67(1.41)	-	-
	Baseline		51.88(0.44)	100	-	0	0
2	AIC	Proposed-I	0.70(0.04)	97.69(0.75)	34.49(15.03)	1.72(0.08)	0.98(0.18)
		Proposed-II	0.71(0.05)	97.86(0.59)	25.12(0.80)	1.72(0.06)	0.94(0.16)
		Proposed-III	0.34(0.02)	98.39(0.96)	73.55(14.28)	1.84(0.05)	2.96(0.20)
		SPAR	5.06(0.44)	95.27(1.30)	106.31(34.62)	1.33(0.09)	0.70(0.16)
	BIC	Proposed-I	1.20(0.07)	95.77(0.78)	48.68(14.63)	1.38(0.09)	0.20(0.06)
		Proposed-II	1.54(0.05)	90.77(1.16)	152.59(34.81)	0.88(0.09)	0.28(0.07)
		Proposed-III	1.24(0.06)	94.53(1.12)	97.38(25.10)	1.22(0.10)	0.34(0.07)
		SPAR	8.12(0.31)	91.92(1.10)	125.98(23.65)	0.83(0.07)	0.20(0.06)
	MW		16.86(0.62)	76.88(1.70)	676.32(76.91)	-	-
	Baseline		0.85(0.01)	100	-	2	0
5	AIC	Proposed-I	0.29(0.02)	93.85(1.54)	165.07(33.78)	4.22(0.12)	1.92(0.23)
		Proposed-II	0.26(0.01)	88.30(2.38)	436.27(114.09)	4.00(0.13)	2.28(0.26)
		Proposed-III	0.16(0.00)	91.04(3.31)	871.54(255.62)	4.40(0.11)	4.72(0.19)
		SPAR	0.17(0.01)	85.88(1.80)	2935.38(160.36)	4.24(0.10)	5.60(0.14)
	BIC	Proposed-I	0.71(0.06)	80.95(1.77)	517.32(78.51)	2.44(0.15)	0.48(0.10)
		Proposed-II	1.30(0.09)	69.04(1.78)	1114.60(107.16)	1.32(0.10)	0.46(0.09)
		Proposed-III	0.86(0.06)	72.92(2.12)	960.43(123.56)	1.72(0.14)	0.72(0.10)
		SPAR	12.67(0.63)	73.88(1.68)	821.46(99.70)	0.82(0.08)	0.26(0.06)
	MW		24.13(0.73)	50.94(1.08)	2465.56(103.93)	-	-
	Baseline		0.34(0.00)	100	-	5	0

To further explore the advantage of the proposed method, we investigate which biases terms are selected in the model. Under the setting (I), the proposed method include $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_3, \boldsymbol{\delta}_6, \boldsymbol{\delta}_{11}$ into the final model. As seen in Figure 4.3(a), there are clear initial bending in the original deformation data since all of the average profile seems to deform in a systematical shape. The initial adjusted bending profile from the proposed method is displayed in Figure 4.3 (b). By adjusting all the profiles by the initial adjustment profile, the adjusted new profiles are displayed in 4.3 (c). Clearly, the adjusted profiles are much smoother compared with initial profiles.

Besides the initial bending term $\boldsymbol{\delta}_0$, the proposed methods also include other adjustment terms for possible bias according to GFoBa procedure. Specifically, under setting (I), $\boldsymbol{\delta}_6$

Table 4.3: Result for Data Generated from the Setting (III)

# of Significant $\boldsymbol{\delta}_k$'s	Criterion	Model	RMSE	$\hat{\text{E}}$	MSE($\hat{\text{E}}$)	TP	FP
0	AIC	Proposed-I	42.20(2.40)	100.04(0.07)	0.30(0.06)	0.00(0.00)	0.14(0.05)
		Proposed-II	47.49(1.74)	99.98(0.10)	0.74(0.21)	0.00(0.00)	0.06(0.03)
		Proposed-III	43.44(2.30)	100.05(0.06)	0.44(0.04)	0.00(0.00)	0.14(0.06)
		SPAR	80.81(2.12)	100.34(0.75)	6.21(0.75)	0.00(0.00)	0.28(0.09)
	BIC	Proposed-I	48.69(0.51)	100.05(0.07)	0.55(0.08)	0.00(0.00)	0.00(0.00)
		Proposed-II	50.48(0.53)	99.99(0.06)	0.24(0.05)	0.00(0.00)	0.00(0.00)
		Proposed-III	49.08(0.60)	99.91(0.07)	0.41(0.05)	0.00(0.00)	0.00(0.00)
		SPAR	56.43(0.62)	100.02(0.24)	3.44(0.45)	0.00(0.00)	0.00(0.00)
	MW		11.78(0.39)	102.16(1.46)	19.84(2.52)	-	-
	Baseline		49.33(0.46)	100	-	0	0
2	AIC	Proposed-I	0.62(0.04)	99.25(0.49)	13.40(5.44)	1.84(0.05)	1.20(0.21)
		Proposed-II	0.67(0.04)	98.57(0.55)	22.49(8.04)	1.76(0.07)	1.02(0.20)
		Proposed-III	0.59(0.03)	99.74(0.31)	9.31(2.31)	1.88(0.05)	0.98(0.17)
		SPAR	3.46(0.40)	93.71(1.74)	189.05(44.91)	1.10(0.10)	1.55(0.20)
	BIC	Proposed-I	1.08(0.07)	96.59(1.03)	63.97(29.94)	1.44(0.09)	0.32(0.09)
		Proposed-II	1.18(0.06)	96.29(0.77)	45.61(12.45)	1.28(0.09)	0.26(0.07)
		Proposed-III	0.81(0.01)	99.90(0.31)	9.04(2.31)	1.88(0.47)	0.82(0.17)
		SPAR	7.91(0.24)	88.55(1.27)	211.41(42.60)	0.65(0.07)	0.55(0.20)
	MW		17.31(0.63)	73.55(1.78)	855.14(79.00)	-	-
	Baseline		0.81(0.01)	100	-	2	0
5	AIC	Proposed-I	0.21(0.01)	89.40(2.24)	365.58(92.25)	4.28(0.10)	2.78(0.28)
		Proposed-II	0.21(0.01)	88.40(2.14)	401.90(74.99)	4.24(0.11)	3.12(0.29)
		Proposed-III	0.19(0.01)	92.19(2.75)	470.30(174.72)	4.32(0.16)	3.52(0.34)
		SPAR	0.23(0.10)	89.33(4.69)	1466.46(339.61)	4.13(0.15)	5.58(0.18)
	BIC	Proposed-I	0.56(0.04)	81.03(1.88)	534.57(79.43)	2.70(0.15)	0.64(0.09)
		Proposed-II	0.73(0.05)	73.72(1.94)	877.90(102.15)	1.90(0.13)	0.76(0.11)
		Proposed-III	0.20(0.01)	92.75(2.57)	414.12(154.69)	4.28(0.16)	3.12(0.31)
		SPAR	1.18(0.07)	70.00(2.09)	1115.22(130.29)	0.75(0.08)	0.38(0.07)
	MW		23.09(0.70)	51.78(0.86)	2362.85(83.42)	-	-
	Baseline		0.32(0.00)	100	-	5	0

Table 4.4: Result for Real Data under AIC as the Model Selection Criterion

Model	RMSE	$1/\hat{\text{E}}$	se($1/\hat{\text{E}}$)	$\hat{E}, GPa$	se($\hat{\text{E}}$)
Proposed-I	0.96	0.98×10^{-2}	6.59×10^{-9}	101.82	0.71
Proposed-II	1.40	0.95×10^{-2}	2.63×10^{-8}	105.77	3.29
Proposed-III	0.81	0.97×10^{-2}	1.96×10^{-8}	103.45	2.25
SPAR	2.20	1.05×10^{-2}	2.05×10^{-8}	94.98	2.53
MW	2.37	1.05×10^{-2}	4.01×10^{-8}	94.96	4.92

entered first, followed by $\boldsymbol{\delta}_3$ and $\boldsymbol{\delta}_{11}$. Under setting (II) and (III) for the covariance structure $\boldsymbol{\Sigma}$, only $\boldsymbol{\delta}_6$ is included as the adjustment term in the model. From Figure 4.3(c), it seems that the original data after adjusted by the initial bending, a downward bias occurs on the experiment when the force level changes to 144nN and affects all the profiles recorded thereafter. Thus

it is reasonable that $\boldsymbol{\delta}_6$ is selected under all the proposed models. Figure 4.3(d) shows the profiles adjusted by the initial bending and the sixth adjustment term. From Figure 4.3(d), we can see that all the profiles with the force level larger than 144nN are adjusted upward to adjust the bias.

4.7 Discussion

In this work, a mixed variance component model is proposed to utilize all available NB profiles data for quantifying the elastic modulus. The mixed variance component model is more accurate in estimation compared with SPAR or MW since it can flexibly accommodate multiple sources of variances. Initial bias and systematic error can occur in the experimental data. A group adaptive forward backward selection procedure is proposed to select the adjustment terms needed where the systematic error occurs. We would like to remark that the mixed variance component model can be applied not only on Nano experiments, but more generally on various studies that contains functional profiles observations too such as the climate regional study (Ramsay, 2006).

For the within profile covariance structure, three variance structures are considered in this paper including the independent setting, AR(1) setting and Gaussian setting. In future work, other commonly used variance covariance settings can also be used such as the compound symmetry setting, Toeplitz setting, etc. (Wolfinger, 1993; Liang and Zeger, 1986).

For the real data analysis in Section 4.6, we have also performed the proposed method under BIC. The result under BIC criterion yields a smaller standard error compared with the corresponding result of AIC but with a larger RMSE. It may be the case that the estimate of E under BIC is more stable but biased compared with that of AIC. Another issue for real data is that outliers could be present in the experimental study. For example, one possible outlier can be observed from the real data in Figure 4.1(a). In future, it would be interesting to develop a functional outlier detection method such that the outliers profile can be removed to further increase the estimation accuracy. One possible solution is to obtain a median curve

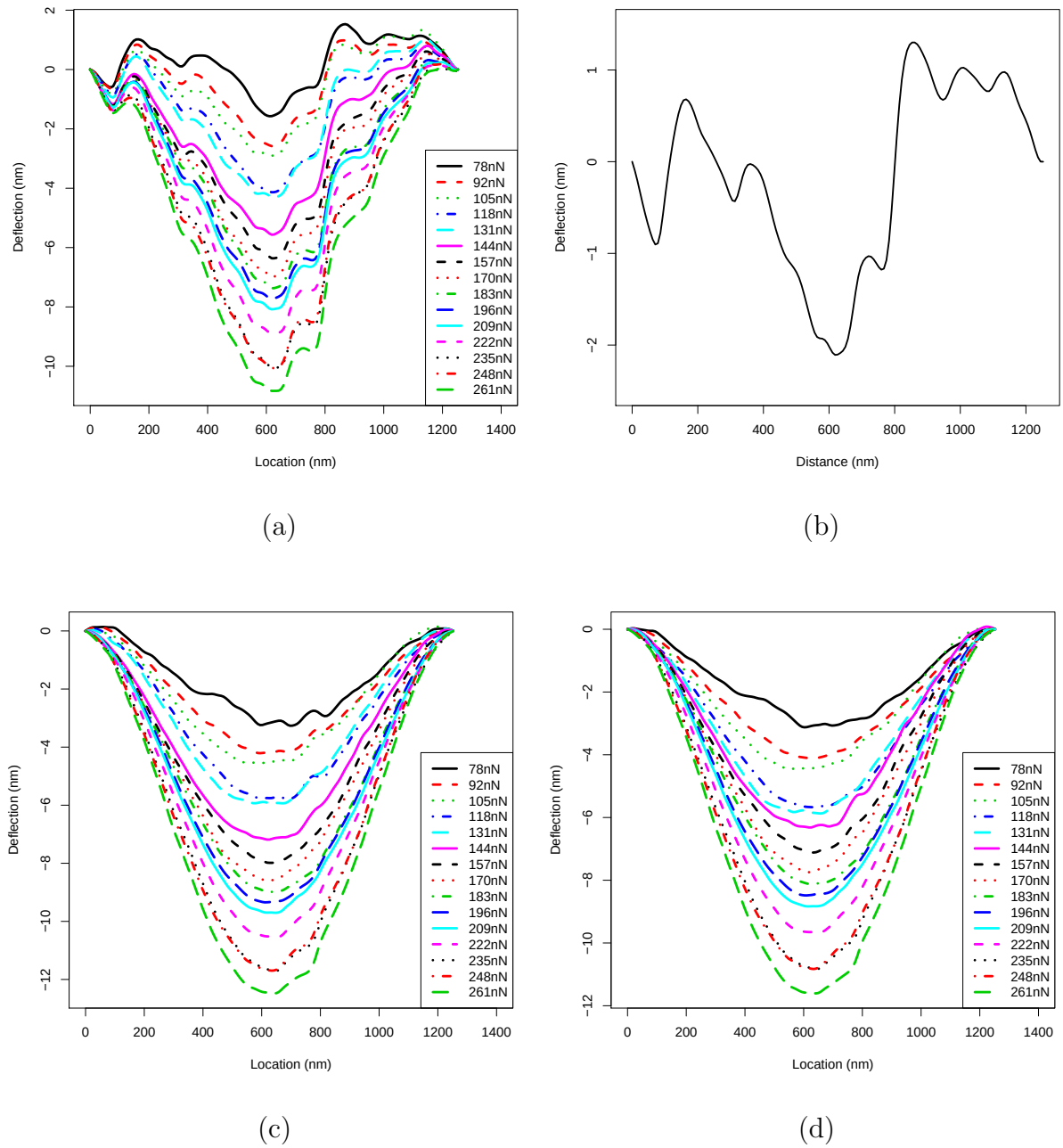


Figure 4.3: (a) Real data with 10 replicates under each of the 15 force levels. (b) The initial adjustment bending profile estimated from the Proposed-I method. (c) The average bending profiles normalized by the initial bending profile. (d) The corresponding profiles of (a) after adjusted by the initial and the sixth adjustment terms under the Proposed-I method.

and detect the curves far away from the median curve as the outliers. Distance measure between two functional data profiles need to be developed (Febrero et al., 2008; Yu et al., 2012).

Chapter 5 Summary

In this thesis, I develop several methods for dealing with data with complex structures under the content of classification and regression modeling. Specifically, a regularized approach to sparse linear discriminant analysis is developed for two-class classification. A two-stage model construction and evaluation for reject inference is developed for accurate and efficient fraud detection. I also developed a mixed variance component model to deal with nano experimental data with multiple sources of variability in the experiments.

For the two-class classification, linear discriminant analysis (LDA) is a commonly used method. However, LDA may not perform well in the case of high dimensional data with number of variables p larger than the number of observations n. In Chapter 2, I proposed a novel sparse LDA which generalizes LDA through a regularized approach with shrinkage on both the inverse covariance matrix and the mean difference between the two classes. The proposed method enjoys the advantage of ease of interpretation combined with efficient computation. Simulation under multiple data generation settings are conducted to compare the proposed method with other commonly used methods in terms of estimation accuracy and misclassification error. The performance of the proposed method is also examined through leukemia and lung cancer real data.

For the reject inference in fraud detection, the training data does not have incomplete information with unknown class labels. Under this context, the usual classifiers are no longer feasible solutions. When the missing labels are missing not at random, there is a need to develop certain calculation schemes to incorporate the information hidden in the observations with missing labels. The proposed two-stage modeling approach is developed to resolve the non-random missing class label issue. Specifically, I studied the reject inference problem in Chapter 3. With the prevalence of the electronic commerce, fraud commonly exists in online business. To prevent fraud transactions, suspicious transactions are rejected with unknown

fraud status by the online decision system. To build an accurate risk model for fraud detection, one great challenge is how to use the information of rejected transactions for modeling and model evaluation. In Chapter 3, I propose a two-stage model to effectively identify known fraud patterns and missing frauds. The proposed model fully exploits the fraud patterns in both accepted and rejected transactions, thus enhances the detection of fraud at both stages. Moreover, I develop a novel model evaluation criterion based on adjusted net profit value, which takes the information on rejected transactions into account. It therefore can help to find the optimal risk model. The performance of the proposed method is illustrated through a real case study of Microsoft Xbox online transaction data.

For the mixed variance component modeling approach to analyzing the nano experiment data, the motivation comes from the fact that the nano data contains multiple sources of variation. As the experiment is conducted at the nano-scale, it is important to accurately characterize the variance of the model such that one can estimate the mechanic property of the nano-material accurately. To achieve this goal, a novel variance component mixed model is employed based on SPAR in Deng et al. (2009). The original nano data are obtained from the ZnO nanobelt experiment, following the physical free-free beam model (FFBM) (Mai and Wang, 2006). Since initial bias and systematic errors may be introduced during the experiment and data collection process, all the profiles data instead of the averaged profile under each experiment group as in Deng et al. (2009) are used in the model to improve the estimation accuracy of the elastic modulus. In Chapter 4, I also develop a group variable selection procedure based on Zhang (2009) as the major variable selection technique to select significant terms that can effectively adjust the systematic errors in the nano data. The performance of the model is illustrated through simulation data that mimic the same data generation scheme of the experiment as well as the real experiment data.

References

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, New York.

Au, W. H., Chan, K. C. C., Wong, A. K. C. and Wang, Y. (2005). Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **2(2)**, 83-101.

Auer, P., Burgsteiner, H. and Maass, W. (2008). A learning rule for very simple universal approximators consisting of a single layer of perceptrons. *Neural Networks*, **21(5)**, 786-795.

Banasik, J. and Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, **183(3)**, 1582-1594.

Benham, P. P., Crawford, R. J. (1987). *Mechanics of Engineering Materials*. John Wiley & Sons, New York.

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

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Annals of Statistics*, **36**, 2577-2604.

Bogner, A., Jouneau, P.H., Thollet, G., Basset, D. and Gauthier, C. (2007). A history of scanning electron microscopy developments: towards "wet-STEM" imaging. *Micron*, **38**, 390-401.

Bolker, B. M., Brooks M. E., Clark C. J., Geange S. W., Poulsen J. R., Stevens M. H. H. and White J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, **24**, 127-135.

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, **17(3)**, 235–255.

Bonate, P. L. (2006). *Pharmacokinetic-Pharmacodynamic Modeling and Simulation*. Springer, New York.

Button, M., Lewis, C. and Tapley, J. (2009). Fraud typologies and the victims of fraud literature review. National Fraud Authority: London.

Cai, T. and Liu, W. (2012). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, **106**, 1566-1577.

Camps-Valls, G., Bandos Marsheva, T. and Zhou, D. (2007). Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*. **45(10)**, 3044-3054.

Cancès, E., Ehrlacher, V. and Lelièvre, T. (2011). Convergence of a greedy algorithm for high-dimensional convex nonlinear problems. *Mathematical Models and Methods in Applied Sciences*, **21(12)**, 2433-2467.

Casella, G. and Berger, R. L. (2002). *Statistical Inference, 2nd edition*. Duxbury press, Pacific Grove, CA.

Chang, C. J. and Joseph, V. R. (2014). Model Calibration Through Minimal Adjustments. *Technometrics*, **56(4)**, 474-482.

Chen, G. and Astebro, T. (2001). The economic value of reject inference in credit scoring. In *Credit Scoring and Credit Control VII: Proceedings of Conference held at University of Edinburgh, Edinburgh, Scotland*.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **95**, 759-771.

Clemmensen, L., Hastie, T., Witten, D. and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, **53**, 406-413.

Copas, J. B. and Li, H. G. (1997). Inference on non-random samples (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59(1)**, 55-95.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20(3)**, 273-297.

Crook, J. and Banasik, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*, **28(4)**, 857-874.

Cui, Y. and Lieber, C. M. (2001). Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science*, **291**, 1289-1292.

Dabney, A. R. (2005). Classification of microarrays to nearest centroids. *Bioinformatics*, **21**, 4148-4154.

Danaher, P., Wang, P. and Daniela, D. M., (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76(2)**, 373-397.

Dasgupta, T., Ma, C., Joseph, V. R., Wang, Z. L. and Wu, C. J. (2008). Statistical modeling and analysis for robust synthesis of nanostructures. *Journal of the American Statistical Association*, **103(482)**, 594-603.

Dasgupta, T., Weintraub, B. and Joseph, V. R. (2011). A physical-statistical model for density control of nanowires. *IIE Transactions*, **43(4)**, 233-241.

Davis, C. (1962). The norm of the Schur product operation. *Numerische Mathematik*, **4(1)**, 343-344.

Davison, A. and Hall, P., (1992). On the bias and variability of bootstrap and cross validation estimates of error rates in discrimination problems. *Biometrika*, **79**, 274-284.

Delamaire, L., Abdou, H. and Pointon, J. (2009). Credit card fraud detection techniques: A review. *Banks and Banks Systems*, **4(2)**, 57-68.

Deng, X., Joseph, V. R., Mai, W., Wang, Z. L., Wu, C. F. J. (2009). Statistical approach to quantifying the elastic deformation of nanomaterials. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 11845-11850.

Djurii, A. B., Choy, W. C. H., Roy, V. A. L., Leung, Y. H., Kwong, C. Y., Cheah, K. W., Rao, T. K. G., Chan, W. K., Lui, H. F. and Surya, C. (2004). Optical properties of ZnO nanostructures. *Advanced Functional Materials*, **14**, 856.

Ethem, A. (2004). *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. MIT Press, Cambridge.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Annals of Statistics*, **36**, 2605-2637.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proceedings International Congress of Mathematicians*, **3**, 595-622.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high-dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70(5)**, 849-911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, **20**, 101-148.

Febrero, M., Galeano, P. and Gonzales-Mantegia, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics*, **19**, 331-345.

Feelders, A. J. (1999). Credit scoring and reject inference with mixture models. *International Journal of Intelligent System in Accounting, Finance and Management*, **8**, 271-279.

Feelders, A. J. (2003). An overview of model based reject inference for credit scoring. *Banff Credit Risk Conference, Banff, Canada*.

Finley, T. and Joachims, T. (2008). Training structural SVMs when exact inference is intractable. *Proceedings of the 25th International Conference on Machine Learning (ICML)*, 304-311.

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

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley-Interscience, Hoboken, NJ.

Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, **23**, 2020-2028.

Freund, Y., Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, **37(3)**, 277-296.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432-441.

Gallant, S. I. (1990). Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks*, **1(2)**, 179-191.

Gao, S., Hendrie, H. C., Hall, K. S. and Hui, S. (1998). The relationships between age, sex, and the incidence of dementia and Alzheimer disease: a meta-analysis. *Archives of General Psychiatry*, **55**, 809-815.

George, A. W., Visscher P. M. and Haley, C. S. (2000). Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, **156**, 2081-2092.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Golub, G. and Van Loan, C. F. (1996). *Matrix Computations (3rd Edition)*. The Johns Hopkins University Press, Baltimore.

Guillaumin, M., Verbeek, J. and Schmid, C. (2010). Multimodal semi-supervised learning for image classification. Proceedings of IEEE Conference Computer Vision and Pattern Recognition, 902-909.

Guo, J., Levina, E., Michailidis, G. and Zhu, J. (2011). Joint estimation of multiple graphical models, *Biometrika*, **98(1)**, 1-15.

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

Hall, P., Titterington, D. M. and Xue, J. H. (2009). Median-based classifiers for high dimensional data. *Journal of the American Statistical Association*, **104**, 1597-1608.

Hand, D. J. and Henley, W. E. (1993). Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry*, **5**, 45-55.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of Royal Statistical Association: Series A (Statistics in Society)*, **160(3)**, 523-541.

Hardin, J. and Hilbe, J. (2007). *Generalized Linear Models and Extensions (2nd edition)*. Stata Press, College Station.

Hastie, T. J., Tibshirani, R. J. and Friedman, J. H. (2008). *Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)*. Springer, Berlin.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153-161.

Hosmer, D. W., Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley & Sons, New York.

Howland, P. and Park, H. (2004). Generalizing discriminant analysis using the generalized singular value decomposition. *IEEE Transactions, Pattern Analysis and Machine Intelligence*, **26(8)**, 995-1006.

Hrushka, D., Kohrt, B. and Worthman, C. A. (2005). Estimating between-and within-individual variation in cortisol using multilevel models. *Psychoneuroendrocrinology*, **30(7)**, 698-714.

Hsia, D. C. (1978). Credit scoring and the equal credit opportunity act. *The Hastings Law Journal*, **30**, 371-448.

Hsu, C. W. and Lin, C. J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions Neural Network*, **13(2)**, 415-425.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, **22(4)**, 679-688.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the sixteenth international conference on machine learning (ICML)*, 200-209.

Joanes, D. N. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Mathematics Applied in Business & Industry*, **5**, 35-43.

Joseph, V. R. and Melkote, S. N. (2009). Statistical adjustments to engineering models. *Journal of Quality Technology*, **41(4)**, 362-375.

Joseph, V. R. and Yan, H. (2014). Engineering-driven statistical adjustment and calibration. *Technometrics*, (just-accepted), 00-00.

Kellam, S., Brown, H. C., Poduska, J., Ialongo, N., Wang, W. and Toyinbo, P. (2008). Effects of a universal classroom behavior management program in first and second grades on young adult behavioral, psychiatric, and social outcomes. *Drug and Alcohol Dependence*, **95**, 5-28.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T. and Xu, Y. (2012). Trustworthy online controlled experiments: Five puzzling outcomes explained. *Proceedings of the 18th Conference on Knowledge Discovery and Data Mining*.

Kotsiantis, S., Zaharakis, I. and Pintelas, P. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, **26(3)**, 159-190.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied linear statistical models, fifth edition*. McGraw-Hill, Irwin.

Laird, N. M. and Ware J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Lachenbruch, P. A. (1979). Discriminant analysis. *Biometrics*, **35**, 69-85.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, **17(4)**, 491-502.

Littell, R. C., Pendergast, J. and Natarajan, R. (2000). Tutorial in biostatistics: modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, **19**, 1793-1819

Mai, W. and Deng, X. (2010). The applications of statistical quantification techniques in nanomechanics and nanoelectronics. *Nanotechnology*, **21(40)**, 405-704.

Mai, W. and Wang, Z. L. (2006). Quantifying the elastic deformation behaviour of bridged nanobelts. *Applied Physics Letters*, **89**, 073-112.

Mai, Q. and Zou, H. (2013). A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics*, **55**, 243-246.

Mai, Q., Zou, H. and Yuan, M., (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, **99**, 29-42.

McCallum, A. and Nigam, K. (1998). A comparison of event models for naive bayes text classification. *Workshop on Learning for Text Categorization*, **752**, 41-48.

McFadden, D. and Train, K. (2000). Mixed MNL models for discrete response. *Journal of Applied Economics*, **15**, 447-470.

McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley & Sons, New York.

Meyer, K. (1999). Estimates of genetic and phenotypic covariance functions for postweaning growth and mature weight of beef cows. *Journal of Animal Breeding and Genetics*, **116**, 181-205.

Montrichard, D. (2008). Reject inference methodologies in credit risk modeling. Paper ST-160, Sesug, Inc.

Morrison, D. G. (1969). On the interpretation of discriminant analysis. *Journal of Marketing Research*, **6**, 156-163.

Nick, L. (1988). Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm. *Machine Learning*, **2**, 285-318.

Ng, A. and Jordan, M. (2002). On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. *Advances in Neural Information Processing Systems (NIPS)*, **14**, 841-848.

Ngai, E. W. T., Xiu, L. and Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature relationship and classification. *Expert Systems with Applications*, **36**, 2529-2602.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y. and Sun, X. (2010). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, **50(3)**, 559-569.

Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39-50.

Nussey, D. H., Wilson, A. J. and Brommer, J. E. (2007). The evolutionary ecology of individual phenotypic plasticity in wild populations. *Journal of Evolutionary Biology*, **20**, 831-844.

Pavlenko, T., Björkström, A. and Tillander, A. (2012). Covariance structure approximation via gLasso in high-dimensional supervised classification. *Journal of Applied Statistics*, **39**, 1643-1666.

Ramsay, J. O. (2006). *Functional data analysis.* John Wiley & Sons, Inc., New York.

Robinson, D. L. (1987). Estimation and use of variance components. *The Statistician*, **36**, 3-14.

Rose, D. and Pevalin, D. J. (2001). *The national statistics socio-economic classification: unifying official and sociological approaches to the conceptualisation and measurement of social class.* ISER Working Paper, 2001-4, Colchester, University of Essex, UK.

Rothman, A., Bickel, P., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, **2**, 494-515.

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

Shaun, S., John, G., Dan, J. and John, C. (2013). What is meant by "missing at random". *Statistical Science*, **28(2)**, 257-268.

Siddiqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. John Wiley & Sons, Hoboken, New Jersey.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A. and Richie, J. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203-209.

Song, J., Xie, H., Wu, W., Joseph, V. R., Wu, C. J. and Wang, Z. L. (2010). Robust optimization of the output voltage of nanogenerators by statistical design of experiments. *Nano Research*, **3(9)**, 613-619.

Stangl, J., Hol, V. and Bauer, G. (2004). Structural properties of self-organized semiconductor nanostructures. *Reviews of Modern Physics*, **76**, 725.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, **50**, 1171-1177.

Puhani, P. A. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, **14**, 53-68.

Takkouche, B., Cadarso-Suarez, C. and Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, **150(2)**, 206-215.

Tapley, B. D., Schutz, B. E. and Born, G. H. (2004). *Statistical Orbit Determination*. Elsevier, Burlington, MA.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **58**, 267-288.

Tibshirani, R., Hastie, T., Narashimhan, B., Chu, G. (2003). Class prediction by nearest shrunken centroids with applications to DNA microarrays. *Statistical Science*, **18**, 104-17.

Tinsley, D. and Stetz, P. E. (2004). Contribution margin pricing for small businesses. *Association of Small Business and Entrepreneurship Conference*.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics*, **2(3)**, 316-325.

Wang, L., Chu, F. and Xie, W. (2007). Accurate cancer classification using expressions of very few genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4(1)**, 40-53.

Wang, F., Hwang, Y., Qian, P. Z. G. and Wang, X. (2010). A statistics-Guided approach to precise characterization of nanowire morphology. *ACS Nano*, **4(2)**, 855-862.

Wang, S. and Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, **23**, 972-979.

Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high dimensional problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71(3)**, 615-636.

Witten, D. M. and Tibshirani, R. J. (2011). Penalized classification using Fishers linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73(5)**, 753-772.

Wolfinger, R.D. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation*, **22(4)**, 1079-1106.

Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C. and Lin, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, **25**, 1145-1151.

Xu, L. and Huang, Q. (2014). Growth process modeling of III-V nanowire synthesis via selective area metal-organic chemical vapor deposition. *IEEE Transactions on Nanotechnology*, **13(6)**, 1093-1101.

Yu, G., Zou, C. and Wang, Z. (2012). Outlier detection in the functional observations with applications to profile monitoring. *Technometrics*, **54**, 308-318.

Yu, M. F., Lourie, O., Dyer, M., Moloni, K., Kelly, T. F. and Ruoff, R. S. (2000). Strength and breaking mechanism of multiwalled carbon nanotubes under tensile load. *Science*, **287**, 637-640.

Yuan, G. X., Ho, C. H. and Lin, C. J. (2011). Recent advances of large-scale linear classification. *Proceedings of the IEEE*, **100(9)**, 2584-2603.

Zhang, T. (2009). Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Advances in Neural Information Processing Systems*, 1921-1928.

Zhu, L., Dasgupta, T. and Huang, Q. (2014). A D-optimal design for estimation of parameters of an exponential-linear grouth curve of nanostructure. *Technometrics*. **56(4)**, 432-442.

Zhu, X., Rogers, T., Qian, R. and Kalish, C. (2007). Humans perform semi-supervised classification too. *Proceedings of the 21st Conference on Artificial Intelligence (AAAI-11)*, 864-870.

Appendix

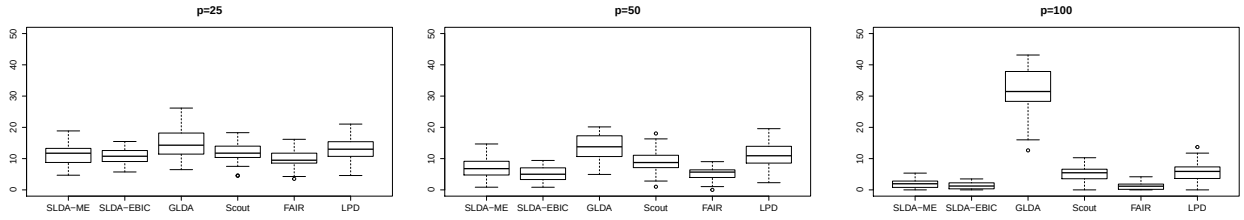
A: Derivation for Objective Function (2.9) (Chapter 2)

When $\boldsymbol{C}$ is set fixed, the objective function (2.9) is equivalent to the objective function of the lasso problem as can be derived as follows:

$$
\begin{aligned}
&\sum_{i\in G_1}(\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}}) \\
&+ \sum_{i\in G_2}(\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}})'\boldsymbol{C}(\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{\delta}_h - \bar{\boldsymbol{x}}) + \lambda_2||\boldsymbol{\delta}_h||_1 \\
=&\sum_{i\in G_1}(\boldsymbol{C}^{1/2}\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{C}^{1/2}\boldsymbol{\delta}_h - \boldsymbol{C}^{1/2}\bar{\boldsymbol{x}})'(\boldsymbol{C}^{1/2}\boldsymbol{x}_i - \frac{2n_2}{n}\boldsymbol{C}^{1/2}\boldsymbol{\delta}_h - \boldsymbol{C}^{1/2}\bar{\boldsymbol{x}}) \\
&+ \sum_{i\in G_2}(\boldsymbol{C}^{1/2}\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{C}^{1/2}\boldsymbol{\delta}_h - \boldsymbol{C}^{1/2}\bar{\boldsymbol{x}})'(\boldsymbol{C}^{1/2}\boldsymbol{x}_i + \frac{2n_1}{n}\boldsymbol{C}^{1/2}\boldsymbol{\delta}_h - \boldsymbol{C}^{1/2}\bar{\boldsymbol{x}}) + \lambda_2||\boldsymbol{\delta}_h||_1 \\
\propto&\sum_{i\in G_1}(-2(\frac{2n_2}{n}\boldsymbol{C}^{1/2}\boldsymbol{\delta}_h)'(\boldsymbol{C}^{1/2}\boldsymbol{x}_i - \boldsymbol{C}^{1/2}\bar{\boldsymbol{x}}) + \frac{4n_2^2}{n^2}\boldsymbol{\delta}_h'\boldsymbol{C}\boldsymbol{\delta}_h) \\
&+ \sum_{i\in G_2}(2(\frac{2n_1}{n}\boldsymbol{C}^{1/2}\boldsymbol{\delta}_h)'(\boldsymbol{C}^{1/2}\boldsymbol{x}_i - \boldsymbol{C}^{1/2}\bar{\boldsymbol{x}}) + \frac{4n_1^2}{n^2}\boldsymbol{\delta}_h'\boldsymbol{C}\boldsymbol{\delta}_h) + \lambda_2||\boldsymbol{\delta}_h||_1 \\
=&-\frac{4n_2}{n}\boldsymbol{\delta}_h'\boldsymbol{C}\sum_{i\in G_1}\boldsymbol{x}_i + \frac{4n_2}{n}\boldsymbol{\delta}_h\boldsymbol{C}(n_1\bar{\boldsymbol{x}}) + \frac{4n_1n_2^2}{n^2}\boldsymbol{\delta}_h'\boldsymbol{C}\boldsymbol{\delta}_h \\
&+ \frac{4n_1}{n}\boldsymbol{\delta}_h'\boldsymbol{C}\sum_{i\in G_2}\boldsymbol{x}_i - \frac{4n_1}{n}\boldsymbol{\delta}_h\boldsymbol{C}(n_2\bar{\boldsymbol{x}}) + \frac{4n_1^2n_2}{n^2}\boldsymbol{\delta}_h'\boldsymbol{C}\boldsymbol{\delta}_h + \lambda_2||\boldsymbol{\delta}_h||_1 \\
=&\frac{4n_1n_2}{n}\boldsymbol{\delta}_h'\boldsymbol{C}\boldsymbol{\delta}_h + \frac{4n_1}{n}\boldsymbol{\delta}_h'\boldsymbol{C}(n\bar{\boldsymbol{x}}) - 4\boldsymbol{\delta}_h'\boldsymbol{C}\sum_{i\in G_1}\boldsymbol{x}_i - \frac{4n_1}{n}\boldsymbol{\delta}_h'\boldsymbol{C}(n\bar{\boldsymbol{x}}) + 4\boldsymbol{\delta}_h'\boldsymbol{C}(n_1\bar{\boldsymbol{x}}) + \lambda_2||\boldsymbol{\delta}_h||_1 \\
=&\frac{4n_1n_2}{n}\boldsymbol{\delta}_h'\boldsymbol{C}\boldsymbol{\delta}_h - 4\boldsymbol{\delta}_h'\boldsymbol{C}(\sum_{i\in G_1}\boldsymbol{x}_i - n_1\bar{\boldsymbol{x}}) + \lambda_2||\boldsymbol{\delta}_h||_1 \\
\propto&\frac{4n_1n_2}{n}(\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\delta}_h)'(\tilde{\boldsymbol{y}} - \tilde{X}\boldsymbol{\delta}_h) + \lambda_2||\boldsymbol{\delta}_h||_1,
\end{aligned}
$$

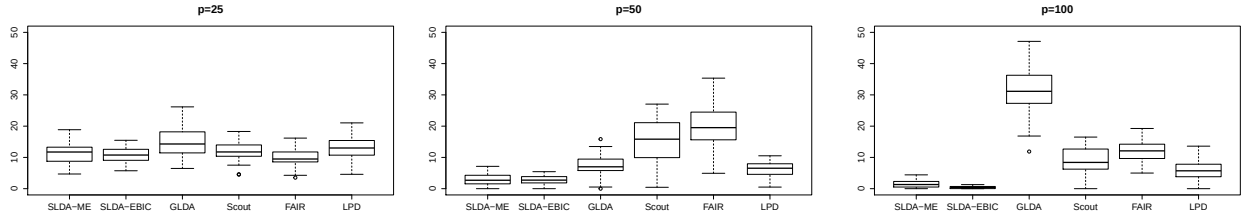
where $\tilde{\boldsymbol{y}} = \frac{n}{2n_1n_2}\boldsymbol{C}^{1/2}(\sum_{i\in G_1}\boldsymbol{x}_i - n_1\bar{\boldsymbol{x}})$, $\tilde{X} = \boldsymbol{C}^{1/2}$.

B: Boxplots of Misclassification Rates for Simulation Study (Chapter 2)



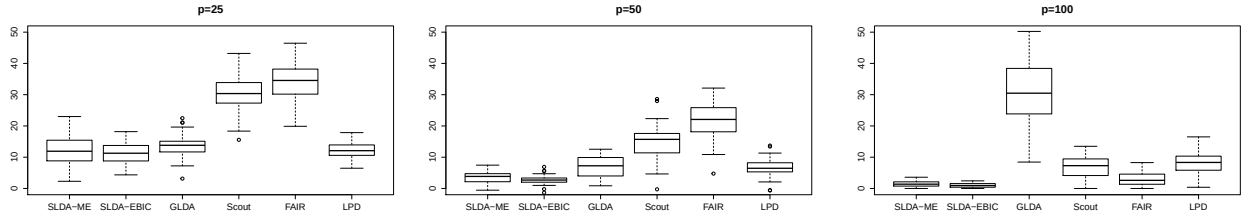
(a) (b) (c)

Figure 1: Misclassification rates from 50 replications under S1 and Model 1 setting for different p: (a) p=25; (b) p=50; (c) p=100.



(a) (b) (c)

Figure 2: Misclassification rates from 50 replications under S1 and Model 2 setting for different p: (a) p=25; (b) p=50; (c) p=100.

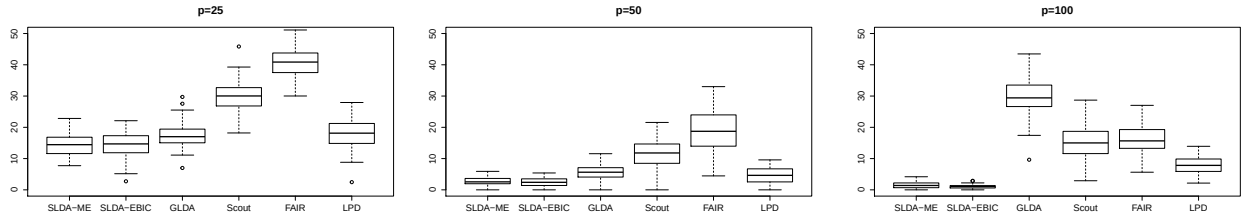


(a) (b) (c)

Figure 3: Misclassification rates from 50 replications under S1 and Model 3 setting for different p: (a) p=25; (b) p=50; (c) p=100.

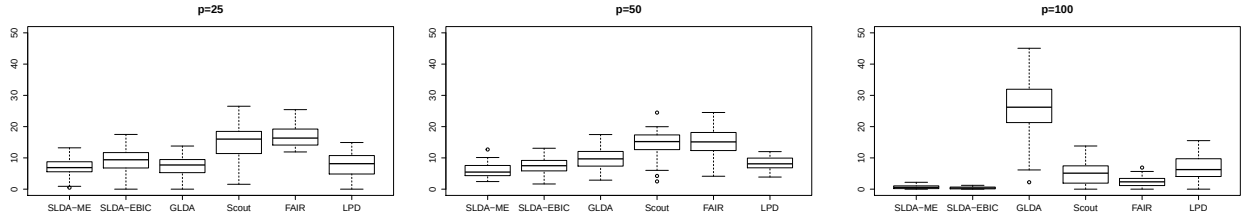
C: Calculation of Total Purchase Amount for Good and Bad Reject Unknown Users (Chapter 3)

Similar as the portfolio decomposition introduced in Section 3, the portfolio of group B data can be decomposed as $S_c = S_{aB} \cup S_{aG} \cup S_{rG} \cup S_{rB}$, where S_{aB} and S_{aG} denote the sets of bad



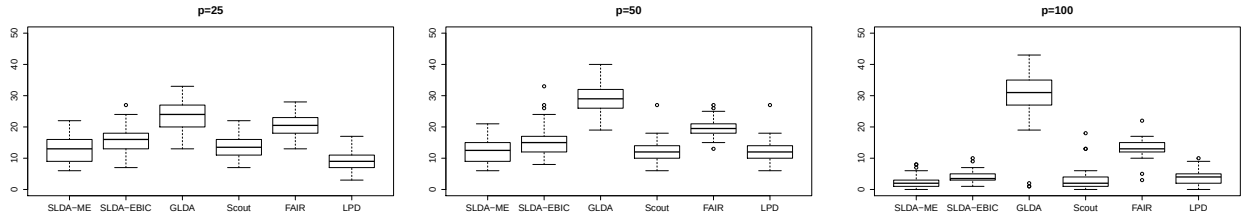
(a) (b) (c)

Figure 4: Misclassification rates from 50 replications under S1 and Model 4 setting for different p: (a) p=25; (b) p=50; (c) p=100.



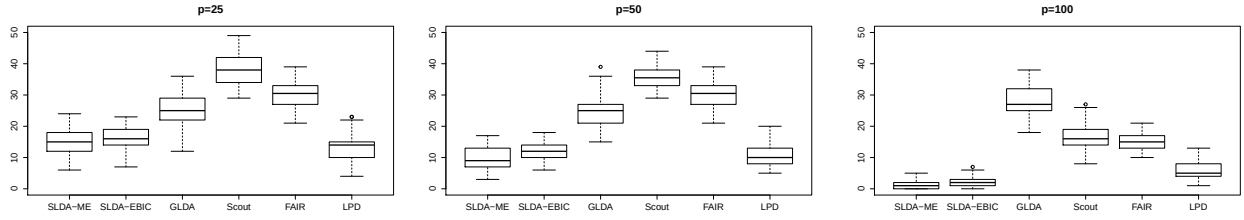
(a) (b) (c)

Figure 5: Misclassification rates from 50 replications under S1 and Model 5 setting for different p: (a) p=25; (b) p=50; (c) p=100.



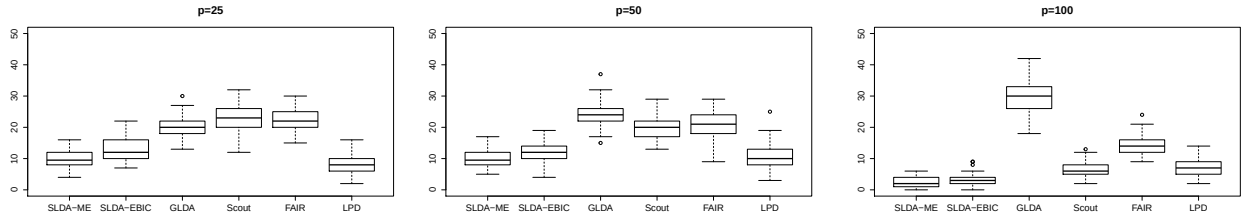
(a) (b) (c)

Figure 6: Misclassification rates from 50 replications under S2 and Model 1 setting for different p: (a) p=25; (b) p=50; (c) p=100.



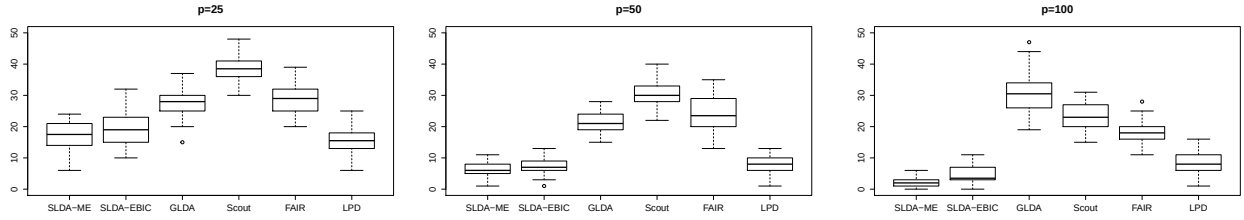
(a) (b) (c)

Figure 7: Misclassification rates from 50 replications under S2 and Model 2 setting for different p: (a) p=25; (b) p=50; (c) p=100.



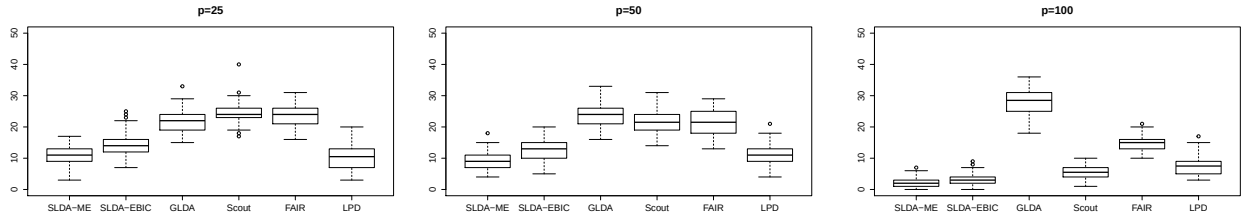
(a) (b) (c)

Figure 8: Misclassification rates from 50 replications under S2 and Model 3 setting for different p: (a) p=25; (b) p=50; (c) p=100.



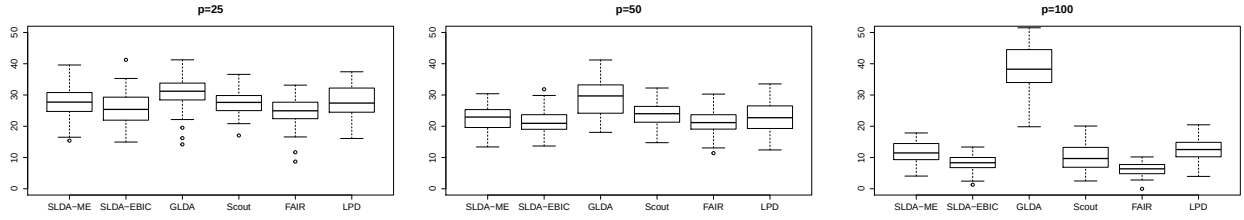
(a) (b) (c)

Figure 9: Misclassification rates from 50 replications under S2 and Model 4 setting for different p: (a) p=25; (b) p=50; (c) p=100.



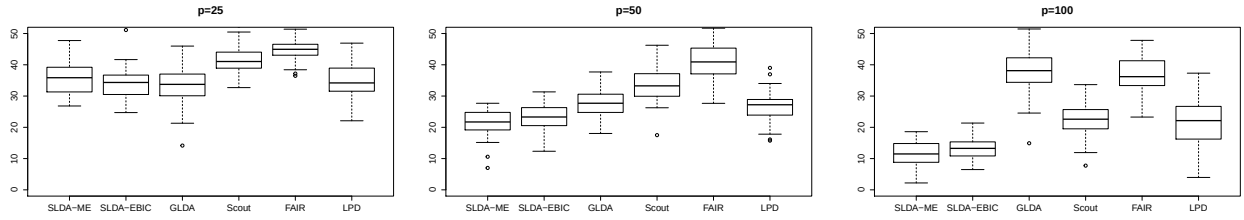
(a) (b) (c)

Figure 10: Misclassification rates from 50 replications under S2 and Model 5 setting for different p: (a) p=25; (b) p=50; (c) p=100.



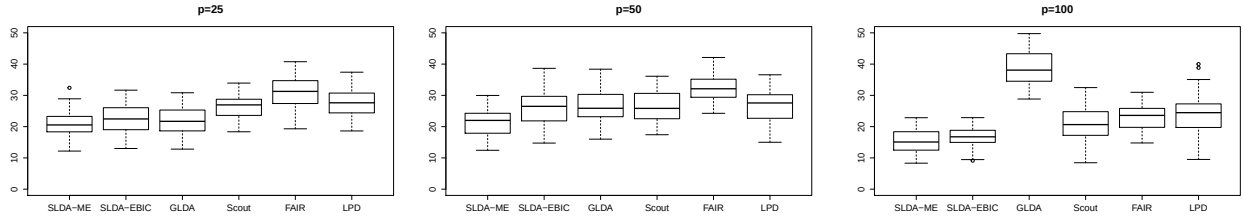
(a) (b) (c)

Figure 11: Misclassification rates from 50 replications under S3 and Model 1 setting for different p: (a) p=25; (b) p=50; (c) p=100.



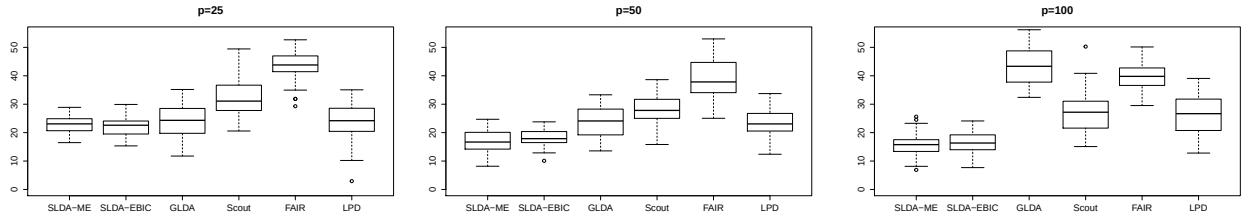
(a) (b) (c)

Figure 12: Misclassification rates from 50 replications under S3 and Model 2 setting for different p: (a) p=25; (b) p=50; (c) p=100.



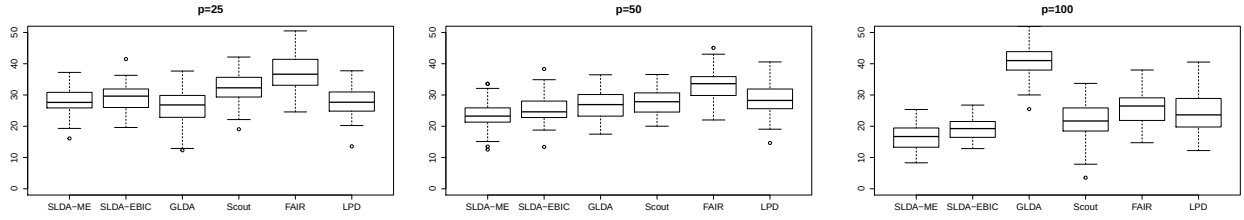
(a) (b) (c)

Figure 13: Misclassification rates from 50 replications under S3 and Model 3 setting for different p: (a) p=25; (b) p=50; (c) p=100.



(a) (b) (c)

Figure 14: Misclassification rates from 50 replications under S3 and Model 4 setting for different p: (a) p=25; (b) p=50; (c) p=100.



(a) (b) (c)

Figure 15: Misclassification rates from 50 replications under S3 and Model 5 setting for different p: (a) p=25; (b) p=50; (c) p=100.

and good transactions respectively with accepted flag while S_{rB} and S_{rG} denote sets of bad and good transactions respectively with rejected flag.

Denote the total purchase amount of transactions in S_{aB}, S_{aG}, S_{rG}, S_{rB} as v_{aB}, v_{aG}, v_{rG}, v_{rB}, respectively and total number of users in S_{aB}, S_{aG}, S_{rG}, S_{rB} as u_{aB}, u_{aG}, u_{rG}, u_{rB}, respectively. Similarly, denote the total purchase amount of transactions in S_{AB}, S_{AG}, S_{RU}, S^G_{RU}, S^B_{RU} as v_{AB}, v_{AG}, v_{RU}, v^G_{RU}, v^B_{RU}, respectively and the total number of users in S_{aB}, S_{aG}, S_{rG}, S_{rB} as u_{aB}, u_{aG}, u_{rG}, u_{rB}, respectively. Denote total purchase amount of $\{t_i \in S_{RG} \text{ and } d(t_i) = 0\}$ as v_{RG} and total purchase amount of $\{t_i \in S_{RB} \text{ and } d(t_i) = 1\}$ as v_{RB}. Assuming the portfolio of rejected transactions in group A is the same as the portfolio of group B transactions with rejected flag, we can calculate the total purchase amount v^B_{RU} and v^G_{RU} as follows. Denote the number of transactions in S_{RU}, S_{rG} as n_{RU}, n_{rG}, respectively. Then the group A users in S_{RU} are rejected $\frac{n_{RU}}{u_{RU}}$ times on average and group B users in S_{rG} perform the purchase action $\frac{n_{rG}}{u_{rG}}$ times on average. Since portfolio of transactions in group A is the same as that of group B, $\frac{n_{rG}}{u_{rG}}$ is also the intended purchase times for S^G_{RU} users. To adjust retrials in rejected transactions, we weight the total purchase amount v^G_{RU} by $w_1 = \frac{n_{rG}}{u_{rG}} / \frac{n_{RU}}{u_{RU}}$. Similarly we can calculate the weight w_2 on v^B_{RU}. We will have:

$$\frac{v_{AG} + v_{RG} + w_1 \times v^G_{RU}}{v_{AB} + v_{RB} + w_2 \times v^B_{RU}} = \frac{v_{aG} + v_{rG}}{v_{aB} + v_{rB}}.$$

The only unknowns in the above equation are v^B_{RU} and v^G_{RU}. Since $v^B_{RU} + v^G_{RU} = v_{RU}$, we can solve for these two numbers.

D: Calculation of Total Number of Good and Bad Users among Reject Unknown Users (Chapter 3)

The total number of good and bad users among reject unknown users can be calculated as follows. Because users in group B is randomly selected, portfolio of transactions in group B should be exactly the same as that in group A. A nature assumption is that the proportion

of good users in group A be the same as that in group B, we have:

$$\frac{u_{AG} + u_{RG} + u_{RU}^{G}}{u_{AB} + u_{RB} + u_{RU}^{B}} = \frac{u_{aG} + u_{rG}}{u_{aB} + u_{rB}}.$$

The only unknowns in the above equation are u_{RU}^{G} and u_{RU}^{B} as a result of unknown of fraud/non-fraud status. Since $u_{RU}^{B} + u_{RU}^{G} = u_{RU}$, we can easily solve for these two numbers.

E: Two-Stage Model with Different Weight Result (Chapter 3)

Table 1 reports result with fixed initial setting $p_1 = 80\%$, $p_2 = 20\%$. The weight on missing fraud is changing from 0 to 35 with step size 1. The bottom line of this table shows result based on the online decision. This table includes result of NPV under group B data and adjusted NPV under validation data. As a comparison, the unadjusted NPV under validation data is calculated using (12) and displayed in the last column. The largest, second largest and third largest NPV under group B data, adjusted NPV under validation data and unadjusted NPV under validation data is colored with red, green and blue respectively.

F: Boxplots for Randomized Data Settings (Chapter 3)

Boxplots from the randomized setting of Section 6.3.4 are displayed here. After modeling using the new training data, I randomly sample 80% of group B data as the new testing data. The random sampling process on group B data is performed 50 times. Three boxplots for NPV, FP and FN results are displayed in Figure 16(a), Figure 16(b) and Figure 16(c), respectively. The results confirm with the findings from the boxplots displayed in Section 6.3.2 based on the original group A data.

Table 1: Two-stage model with different weight result under $p_1 = 0.8, p_2 = 0.2$

Weight	**Under Group B Data**	**Under Validation Data**	
z	NPV	Adjusted NPV	Unadjusted NPV
1	84676.47	228741.46	141235.14
2	84917.85	228805.99	141409.79
3	84810.01	229083.28	141668.81
4	84979.67	229329.54	141856.20
5	84941.93	229507.08	142090.59
6	84981.94	229557.75	142045.15
7	85138.12	229864.55	142279.53
8	85074.73	229636.81	141938.17
9	85040.24	229768.05	142066.76
10	85147.70	229908.14	143098.77 (1)
11	85248.40	229905.40	141884.90
12	85284.99 (3)	230074.38 (2)	142451.01
13	85286.69 (1)	230102.83 (3)	142500.79
14	85256.34 (2)	230151.82 (1)	142611.21
15	85185.27	229949.74	142437.03
16	85161.14	229885.57	142520.11
17	85172.34	229878.71	142526.95
18	85208.36	229915.80	142575.57
19	85194.43	229876.30	142362.37
20	85103.45	229820.63	142318.91
21	85128.08	229919.59	142399.53
22	85160.37	229993.88	142569.56
23	85176.45	230013.84	142552.15
24	85125.71	229884.82	142430.29
25	85203.39	230072.65	142576.33
26	85208.38	230138.33	142687.20 (2)
27	85146.89	229849.54	142624.09
28	85146.89	229872.33	142646.22
29	85174.77	229876.98	142682.36
30	85184.83	229927.90	142717.56 (3)
31	85168.91	229877.62	142649.44
32	85186.24	229894.02	142641.02
33	85211.96	229863.66	142567.25
34	85211.96	229873.63	142629.18
35	85194.64	229894.32	142568.77
36	85126.33	229674.57	142379.12
37	85130.33	229619.09	142336.18
38	85138.33	229647.55	142356.94
39	85147.05	229668.91	142412.64
40	85147.05	229697.53	142441.26
Online Decision	66479.43	210564.41	210564.41

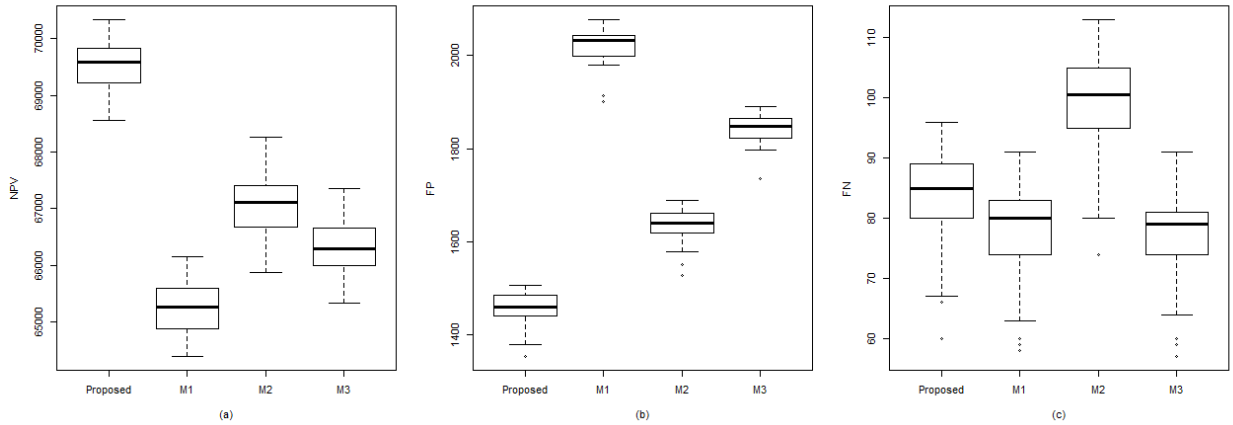


Figure 16: Boxplots under randomized group A data. (a) NPV result under randomized group B data; (b) FP result under randomized group B data; (c) FN result under randomized group B data

G : Prove of Theorem 1 (Chapter 4)

(i). Following the notations in Section 4.4, to prove $||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \leq g_1(k,n)$, it is equivalent to finding an upper bound for the left hand side of the inequality. We have

$$
\begin{aligned}
&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
=&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 + ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
&- ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}E\boldsymbol{y}||_2^2 + ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2.
\end{aligned}
$$

First, there exist upper bound for $||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2$ and similarly for $||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2$. Applying the oracle inequality result in *Theorem 3.1* in Zhang (2011), there exist an upper bound $q(k,n)$ for $||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\beta} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2$ under certain conditions, then an upper bound can be formed for GFoBa under the mixed variance component model.

Denote $\boldsymbol{\Theta} = \boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y}$, write $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ as $\boldsymbol{\Sigma}^{-\frac{1}{2}} = (\boldsymbol{\sigma}_1', \cdots, \boldsymbol{\sigma}_p')$ and $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ as $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} = (\hat{\boldsymbol{\sigma}}_1', \cdots, \hat{\boldsymbol{\sigma}}_p')$, where each $\boldsymbol{\sigma}_i$ and $\hat{\boldsymbol{\sigma}}_i$ is a vector of the ith column of $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$, respectively. Similarly, $\boldsymbol{\Theta}$ can be written by its columns as $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_p)$. Stretching out the terms inside $\boldsymbol{\Theta}, \boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ using their column vectors, we will have:

$$
\begin{aligned}
&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
=&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Theta}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Theta}||_2^2 \\
=&\sum_{i=1}^{p}\sum_{j=1}^{p}(\hat{\boldsymbol{\sigma}}_i'\boldsymbol{\theta}_j)^2 - \sum_{i=1}^{p}\sum_{j=1}^{p}(\boldsymbol{\sigma}_i'\boldsymbol{\theta}_j)^2 \\
=&(\hat{\boldsymbol{\sigma}}_1'\boldsymbol{\theta}_1)^2 + \cdots + (\hat{\boldsymbol{\sigma}}_1'\boldsymbol{\theta}_p)^2 + (\hat{\boldsymbol{\sigma}}_2'\boldsymbol{\theta}_1)^2 + \cdots + (\hat{\boldsymbol{\sigma}}_2'\boldsymbol{\theta}_p)^2 + \cdots + (\hat{\boldsymbol{\sigma}}_p'\boldsymbol{\theta}_p)^2 \\
&- (\boldsymbol{\sigma}_1'\boldsymbol{\theta}_1)^2 - \cdots - (\boldsymbol{\sigma}_1'\boldsymbol{\theta}_p)^2 - (\boldsymbol{\sigma}_2'\boldsymbol{\theta}_1)^2 - \cdots - (\boldsymbol{\sigma}_2'\boldsymbol{\theta}_p)^2 - \cdots - (\boldsymbol{\sigma}_p'\boldsymbol{\theta}_p)^2 \\
=&(\hat{\sigma}_{11}\theta_{11} + \cdots + \hat{\sigma}_{1p}\theta_{p1})^2 - (\sigma_{11}\theta_{11} + \cdots + \sigma_{1p}\theta_{p1})^2 + (\hat{\sigma}_{11}\theta_{12} + \cdots + \hat{\sigma}_{1p}\theta_{p2})^2 \\
&- (\sigma_{11}\theta_{12} + \cdots + \sigma_{1p}\theta_{p2})^2 + \cdots + (\hat{\sigma}_{p1}\theta_{1p} + \cdots + \hat{\sigma}_{pp}\theta_{pp})^2 - (\sigma_{p1}\theta_{1p} + \cdots + \sigma_{pp}\theta_{pp})^2
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{p} (\hat{\sigma}_{1i}^2 - \sigma_{1i}^2)\theta_{i1}^2 + 2\sum_{i\neq j} (\hat{\sigma}_{1i}\hat{\sigma}_{1j} - \sigma_{1i}\sigma_{1j})\theta_{i1}\theta_{j1} + \cdots + \sum_{i=1}^{p} (\hat{\sigma}_{pi}^2 - \sigma_{pi}^2)\theta_{ip}^2 \\
&\quad + 2\sum_{i\neq j} (\hat{\sigma}_{pi}\hat{\sigma}_{pj} - \sigma_{pi}\sigma_{pj})\theta_{ip}\theta_{jp} \\
&= \sum_{j}\sum_{i} (\hat{\sigma}_{ji}^2 - \sigma_{ji}^2)\theta_{ij}^2 + 2\sum_{k}\sum_{i\neq j} (\hat{\sigma}_{ki}\hat{\sigma}_{kj} - \sigma_{ki}\sigma_{kj})\theta_{ik}\theta_{jk}.
\end{aligned}
$$

Since

$$
\hat{\sigma}_{ki}\hat{\sigma}_{kj} - \sigma_{ki}\sigma_{kj} = \hat{\sigma}_{ki}\hat{\sigma}_{kj} - \hat{\sigma}_{ki}\sigma_{kj} + \hat{\sigma}_{ki}\sigma_{kj} - \sigma_{ki}\sigma_{kj} = \hat{\sigma}_{ki}(\hat{\sigma}_{kj} - \sigma_{kj}) + \sigma_{kj}(\hat{\sigma}_{ki} - \sigma_{ki}). \quad (1)
$$

Plugging in (1), we will have:

$$
\begin{aligned}
&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
=&\sum_{j}\sum_{i} (\hat{\sigma}_{ji}^2 - \sigma_{ji}^2)\theta_{ij}^2 + 2\sum_{k}\sum_{i\neq j} (\hat{\sigma}_{ki}\hat{\sigma}_{kj} - \sigma_{ki}\sigma_{kj})\theta_{ik}\theta_{jk} \\
=&(\hat{\sigma}_{11}^2 - \sigma_{11}^2)\theta_{11}^2 + 2\hat{\sigma}_{11}(\hat{\sigma}_{12} - \sigma_{12})\theta_{11}\theta_{21} + 2\sigma_{12}(\hat{\sigma}_{11} - \sigma_{11})\theta_{11}\theta_{21} + \\
&(\hat{\sigma}_{12}^2 - \sigma_{12}^2)\theta_{21}^2 + \cdots + (\hat{\sigma}_{pp}^2 - \sigma_{pp}^2)\theta_{pp}^2.
\end{aligned}
$$

Let $\boldsymbol{\Delta} = \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}$, write $\boldsymbol{\Delta}$ as $(\boldsymbol{\delta}_1', \cdots, \boldsymbol{\delta}_p')$. We have:

$$
\begin{aligned}
&||(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}})\boldsymbol{\Theta}||_2^2 \\
=&||\boldsymbol{\Delta}\boldsymbol{\Theta}||_2^2 \\
=&[(\hat{\boldsymbol{\sigma}}_1 - \boldsymbol{\sigma}_1)'\boldsymbol{\theta}_1]^2 + \cdots + [(\hat{\boldsymbol{\sigma}}_1 - \boldsymbol{\sigma}_1)'\boldsymbol{\theta}_p]^2 + \cdots + [(\hat{\boldsymbol{\sigma}}_p - \boldsymbol{\sigma}_p)^T\boldsymbol{\theta}_p]^2 \\
=&[(\hat{\sigma}_{11} - \sigma_{11})\theta_{11} + \cdots + (\hat{\sigma}_{1p} - \sigma_{1p})\theta_{p1}]^2 + \cdots + [(\hat{\sigma}_{p1} - \sigma_{p1})\theta_{1p} + \cdots + (\hat{\sigma}_{pp} - \sigma_{pp})\theta_{pp}]^2 \\
=&\hat{\sigma}_{11}^2\theta_{11}^2 - 2\hat{\sigma}_{11}\sigma_{11}\theta_{11}^2 + \sigma_{11}^2\theta_{11}^2 + 2\hat{\sigma}_{11}\hat{\sigma}_{12}\theta_{11}\theta_{21} - 2\hat{\sigma}_{11}\sigma_{12}\theta_{11}\theta_{21} - 2\sigma_{11}\hat{\sigma}_{12}\theta_{11}\theta_{21} + \\
&2\sigma_{11}\sigma_{12}\theta_{11}\theta_{21} + \hat{\sigma}_{12}^2\theta_{21}^2 - 2\hat{\sigma}_{12}\sigma_{12}\theta_{21}^2 + \sigma_{12}^2\theta_{21}^2 + \cdots + \hat{\sigma}_{pp}^2\theta_{pp}^2 - 2\hat{\sigma}_{pp}\sigma_{pp}\theta_{pp}^2 + \sigma_{pp}^2\theta_{pp}^2.
\end{aligned}
$$

It is straight forward to derive the deduction of the two terms as follows:

$$
\begin{aligned}
&||(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}})\boldsymbol{\Theta}||_2^2 - (||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Theta}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Theta}||_2^2) \\
=&2\sigma_{11}^2\theta_{11}^2 - 2\hat{\sigma}_{11}\sigma_{11}\theta_{11}^2 + 2\sigma_{11}\sigma_{12}\theta_{12}\theta_{21} + 2\sigma_{11}\sigma_{12}\theta_{11}\theta_{21} - 2\sigma_{11}\hat{\sigma}_{12}\theta_{11}\theta_{21} \\
&- 2\hat{\sigma}_{11}\sigma_{12}\theta_{11}\theta_{21} + 2\sigma_{12}^2\theta_{21}^2 - 2\hat{\sigma}_{12}\sigma_{12}\theta_{21}^2 + \cdots + 2\sigma_{pp}^2\theta_{pp}^2 \\
=&\sigma_{11}\theta_{11}^2(\sigma_{11} - \hat{\sigma}_{11}) + \theta_{11}\theta_{21}(\sigma_{12} - \hat{\sigma}_{12})\sigma_{11} + \theta_{11}\theta_{21}(\sigma_{11} - \hat{\sigma}_{11})\sigma_{12} + \\
&+ \sigma_{12}\theta_{21}^2(\sigma_{12} - \hat{\sigma}_{12}) + \cdots + \sigma_{pp}\theta_{pp}^2(\sigma_{pp} - \hat{\sigma}_{pp}) \\
=&\boldsymbol{\sigma}_1'\boldsymbol{\theta}_1\boldsymbol{\delta}_1'\boldsymbol{\theta}_1 + \cdots + \boldsymbol{\sigma}_1'\boldsymbol{\theta}_p\boldsymbol{\delta}_1'\boldsymbol{\theta}_p + \boldsymbol{\sigma}_2'\boldsymbol{\theta}_1\boldsymbol{\delta}_2'\boldsymbol{\theta}_1 + \cdots + \boldsymbol{\sigma}_1'\boldsymbol{\theta}_p\boldsymbol{\delta}_1'\boldsymbol{\theta}_p + \cdots + \boldsymbol{\sigma}_p'\boldsymbol{\theta}_p\boldsymbol{\delta}_p'\boldsymbol{\theta}_p \\
=&|\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Theta} \circ (\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}})\boldsymbol{\Theta}|_1,
\end{aligned}
$$

where $\circ$ denotes the Schur product (Davis, 1962). After some algebra we have $||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{\Theta}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Theta}||_2^2 = ||(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}})\boldsymbol{\Theta}||_2^2 - |\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Theta} \circ (\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}})\boldsymbol{\Theta}|$. So an upper bound can be derived for $||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2$ as follows:

$$
\begin{aligned}
&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
=&(||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2) + (||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
&- ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}E\boldsymbol{y}||_2^2) + (||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2) \\
=&||(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}})(\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y})||_2^2 - ||(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}})(\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y})||_2^2 - |\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y}) \circ \\
&(\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}})(\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y})| + |\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y}) \circ (\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}})(\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y})| + \\
&(||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2).
\end{aligned}
$$

Assume that the covariates are all normalized, from theorem 3.1 in Zhang (2009), under assumption (A2) and (A3), the standardized $\boldsymbol{X}$ follows $||\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y}||_2^2 \leq q(k,n)$ such that when $\boldsymbol{y} \sim N(\boldsymbol{X}\bar{\boldsymbol{\beta}}, \boldsymbol{\Sigma})$, $||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \leq q(k,n)$. Note that assumption (A2) and (A3) is the same assumption as stated in Theorem 3.1 in

Zhang (2009). Then it can be derived that

$$
\begin{aligned}
&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
\leq &||(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}})(\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y})||_2^2 - ||(\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}} - \boldsymbol{\Sigma}^{-\frac{1}{2}})(\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y})||_2^2 - |\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y}) \circ \\
&(\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}})(\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y})| + |\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y}) \circ (\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}})(\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y})| + q(k,p).
\end{aligned}
$$

Given the estimated $\hat{\boldsymbol{\Sigma}}$ is the maximum likelihood estimate of the true $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ will also be the maximum likelihood estimate of $\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}$ by the invariant property with respect to transformation. Maximum likelihood estimate enjoys the asymptotic property such that $\lim_{n\to\infty} ||\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}||_2^2 \to 0$. Assume that $||\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}||_2^2 \leq c$ where c is a fixed small value, we have

$$
\begin{aligned}
&||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 \\
\leq &c(||\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y}||_2^2 - |\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y}) \circ (\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \mathrm{E}\boldsymbol{y})| + |\boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X}\bar{\boldsymbol{\beta}} - \\
&\mathrm{E}\boldsymbol{y}) \circ (\boldsymbol{X}\bar{\boldsymbol{\beta}} - \mathrm{E}\boldsymbol{y})|) + q(k,n) \\
\triangleq &g_1(k,n),
\end{aligned}
$$

where $\lim_{n\to\infty} g_1(k,n) \to 0$. Thus (a) is proved. □

(ii). It has been proved in (a) that $||\hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 - ||\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta}^{(k)} - \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathrm{E}\boldsymbol{y}||_2^2 = |\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Theta} \circ (\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}})\boldsymbol{\Theta}|_1 \triangleq g_2(k,n)$. For the same reasoning as in (a), $\lim_{n\to\infty} ||\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}||_2^2 \to 0$. Assume that $||\boldsymbol{\Sigma}^{-\frac{1}{2}} - \hat{\boldsymbol{\Sigma}}^{-\frac{1}{2}}||_2^2 \leq c$ where c is a fixed small value, we have $\lim_{n\to\infty} g_2(k,n) \to 0$.□