Shrinkage Estimators for Mixture of Linear and Logistic Regression Models

Authored by

© Elsayed Yehia Ghanem

Supervised by

Professor Armin Hatefi
Professor Hamid Usefi

A Thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science.

Department of Mathematics and Statistics
Memorial University of Newfoundland

November 2022

St. John's, Newfoundland and Labrador, Canada

Abstract

The mixture of regression models is one of the most common model-based techniques to incorporate the information of covariates into learning population heterogeneity. The multicollinearity problem is one of the most common problems in regression and a mixture of regression models where the covariates are highly correlated. This problem results in unreliable maximum likelihood estimates for the regression coefficients. In the first part of this thesis, we developed two shrinkage methods through an unsupervised learning approach to estimate the model coefficients in the presence of multicollinearity issues. These shrinkage methods include Ridge and Liu-type estimators. The estimation and prediction performance of the methods are evaluated via EM algorithms.

In the second part of the thesis, we focus on extending the mixture analysis to the binary response in the presence of multicollinearity. The logistic regression model is one of the most powerful statistical methods for analysis of binary data. The logistic regression allows to use a set of covariates to explain the binary responses. The mixture of logistic regression models is used to fit heterogeneous populations through an unsupervised learning approach. This research developed Ridge and Liu-type shrinkage methods to deal with the multicollinearity in a mixture of logistic regression models.

Through extensive numerical studies, we show that the developed methods provide more reliable results in estimating the coefficients of the mixture models. We applied the shrinkage methods to analyze the bone disorder status of women aged 50 and older.

This work is dedicated to my family.

Acknowledgments

First of all, give thanks to Allah, pray in all day full for what is and what was, and if I should count the favors of Allah, I couldn't enumerate them. Indeed, Allah is merciful. I want to focus on people who have supported me and given their hands to me. I would like to thank my supervisors, Dr. Armin Hatefi and Dr. Hamid Usefi, for suggesting the research points and their constant encouragement, leading, friendliness, unlimited support, and giving me valuable comments throughout these years at the Memorial University of Newfoundland. I gratefully admit the financial support that has been provided by Memorial University of Newfoundland School of Graduate Studies, the Department of Mathematics and Statistics, and my supervisors.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This thesis focuses on the finite mixture of linear regression models and finite mixture of logistic regression models. The maximum likelihood (ML) estimate is one of the most common methods for fitting linear and logistic regression models. Although ML estimates are common, the ML estimates are significantly unstable in the presence of multicollinearity, where the regression covariates linearly depend on each other.

Similar to the linear regression and logistic regression models, multicollinearity significantly impacts the ML estimates of the mixture of logistic and mixture of linear regression models. We develop the Liu-type (LT) shrinkage estimation method for the mixture of linear regression models and the mixture of logistic regression models. We show that the LT estimators outperform their Ridge and ML counterparts in estimating the parameters of mixture of linear regressions and mixture of logistic regressions through various simulations and real data studies.

This chapter is organized as follows. Section 1.1 gives an introduction to finite mixture models. Section 1.2 discusses the literature review about the multicollinear-

ity problem. Section 1.3 discusses penalized maximum likelihood estimation. Finally, Section 1.4 presents a literature review of the mixture regression models and mixture of logistic regression models.

1.1 Finite mixture models

Finite mixture models (FMMs) are powerful and practical tools for mathematically modeling populations with multiple subpopulations. Recently, finite mixture models have become more popular for analyzing complex data. Because of their flexibility, mixture models can describe various random events. As a result, they represent complicated processes and systems in many domains of study, including clustering, density estimation, and classification.

Suppose X is a continuous random variable that represents the study population and follows a finite mixture model with M subpopulations. The probability density function (pdf) of the random variable X is given by

$$f(x, \mathbf{\Psi}) = \sum_{j=1}^{M} \pi_j f_j(x, \beta_j), \tag{1.1}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ represents the vector of the mixing proportions with $\pi_j > 0$ and $\sum_{j=1}^{M} \pi_j = 1$ and f_j, $j = 1, \ldots, M$ represents the pdf of the jth component of the model. We use $\mathbf{\Psi} = (\boldsymbol{\pi}, \beta)$ where $\beta^\top = (\beta_1^\top, \beta_2^\top, \ldots, \beta_M^\top)$ represents the vector of all unknown component parameters.

Finite mixture models have been employed in astronomy, biology, genetics, medicine, psychiatry, economics, engineering, marketing, and many other biological,

physical, and social sciences. For example, sodium and lithium counter-transport (SLC) activity in red blood cells is important in quantitative genetics. Furthermore, SLC activity is easier to investigate than blood pressure. Assume that the action of a specific gene specifies the SLC with alleles A and a. The presence of a relevant gene was evaluated utilizing FMMs for analysis of the SLC groups by Chen et al. (2012). The FMMs have also been used in genetics (Schork et al., 1996; Roeder, 1994 and Chen and Chen, 2003), medical studies (Schlattmann, 2009) and various engineering fields, including speech recognition and medical imaging (El Zaart et al., 2002).

Pearson (1894) and Cohen (1967) utilized the method of moments to estimate the parameters in the finite mixture models. Harding (1949) and Cassie (1954) used graphical methods to estimate the finite mixture models. Maximum likelihood (ML) estimation is the most frequent method for estimating the parameters of the mixture models among all methods (Furman and Lindsay, 1994). Here we use EM-algorithm (Dempster et al., 1977) to obtain ML estimates of the parameters of the FMMs.

1.2 Multicollinearity

Regression model is one of the most important methods to explain the relationship between variables. These variables are called explanatory and response variables. The regression model is one common application that comes to mind when predicting response based on a set of explanatory variables (Navidi, 2011). Selecting an appropriate model type based on the characteristics of the result variable, choosing

the explanatory (independent) variables to include in a model, and planning and carrying out model diagnostics are all parts of designing regression models (Shmueli, 2010).

The logistic regression model is one of the most important statistical methods to predict the outcome for a binary response Y based on a set of p covariates $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$. According to Kain and Verma (2018), Logit is one of the most common link functions used in a logistic regression model. The Logit is a function that maps probability values from $(0, 1)$ into real numbers in $(-\infty, +\infty)$. Some distribution functions have been suggested for analyzing a dichotomous (binary) outcome variable (Cox and Snell, 1989). Logistic regression has many applications in various medical research and natural sciences fields. For example, Boyd et al. (1987) used logistic regression to predict mortality in injured patients.

Multicollinearity is a statistical phenomenon that happens when there are high linear dependencies among the independent variables. This problem frequently happens when the model contains many covariates. The primary point is that obtaining accurate estimates of regression coefficients of two or more variables' impact on a particular dependent variable is challenging when they strongly correlate.

According to Lafi and Kaneene (1992), multicollinearity has some primary symptoms: the coefficients' estimate has a high variance, the sign of a coefficient's variable can be different from the theory, and a high correlation between independent variables and outcome. Multicollinearity is a significant issue if the simple correlation coefficient between two regressors is more than 0.8 or 0.9 (Mason and Perreault Jr, 1991). Multicollinearity impacts estimating the coefficients of specific predictors; it has no impact on the predictive accuracy or reliability of the model as a whole

(Mayers, 1990). In statistical modeling, multicollinearity has historically been seen as a huge monster. One of the most challenging tasks in studying statistical modeling has been taming this monster. In regression modeling, the multicollinearity problem is the leading cause for concern among researchers. As a result, multicollinearity causes the variances of parameter estimates to increase. It can also lead to inaccurate estimations of the signs of the regression coefficients, which leads to incorrect inferences about the relationships between explanatory and response variables (Kutner et al., 2004). High multicollinearity causes the confidence intervals of the coefficients to become very wide. When multicollinearity is present, it is thus challenging to reject the null hypothesis of any study (Allison, 1999). To solve this problem, Hoerl and Kennard (1970) proposed ridge regression estimate $\widehat{\boldsymbol{\beta}}_R = \left(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$, where $\boldsymbol{X}$ is a known $n \times p$ design matrix of covariate values, where $n > p$, $\boldsymbol{y}$ is a $n \times 1$ response vector, k is a tunning parameter and $\boldsymbol{I}_{pp}$ is an identity matrix of size $p \times p$. This method has become one of the most popular methods to overcome the weakness of least squares estimators. Ridge estimators attempt to solve the collinearity problem by adding a small constant k to the diagonal of $\boldsymbol{X}^\top \boldsymbol{X}$.

1.3 Penalized Maximum Likelihood Estimation

Penalized maximum likelihood estimation (PMLE) has been proposed to avoid estimation problems, mainly when the likelihood is flat, making a determination of the maximum likelihood (ML) estimate is difficult using standard approaches. It has been successfully applied to stabilize parameter estimates in various models,

such as logistic regression (Firth, 1993; Heinze and Schemper, 2002), latent class models (DeCarlo, 2012). Many improvements have already been made using penalized variable selection methods. Donoho and Johnstone (1994) and Tibshirani (1996) presented the L_1 penalty $p_k(|\beta|) = k|\beta|$, which produces the soft threshold rule. Segerstedt (1992) expanded on Hoerl and Kennard (1970) discussion of the L_2 penalty $p_k(|\beta|) = k|\beta|^2$ results in a Ridge regression, where $p_k(|\cdot|)$ is the penalty function for the coefficients. Antoniadis (1997) studied the hard thresholding penalty function that leads to the hard thresholding rule.

Maximizing the log-likelihood in regular regression modeling provides the best fit for the data set. However, maximizing the log-likelihood frequently leads to fitting noise and unstable parameter estimations when the data set is small; this is due to maximum likelihood estimation placing too much trust in the frequently restricted data trends. The PMLE is developed for regression models and is a generalization of the Ridge regression method used to obtain more stable parameters for linear regression models (Draper and Smith, 1998). The PMLE maximizes the penalized log-likelihood rather than the log-likelihood, where a penalty factor k adjusts the maximum log-likelihood of the model:

$$\log L - 0.5k \sum P(\boldsymbol{\beta}),$$

where L is the maximum likelihood of the fitted model, k is a penalty factor, $\boldsymbol{\beta}$ is a vector of the regression coefficients (Harrell Jr et al., 1998; Van Houwelingen, 2001).

Ridge regression is an ordinary least squares (OLS) estimation with restrictions on the sum of the squared coefficients. Ridge regression may overcome the multicollinearity problem when the linear dependencies between the independent

variables is not severe. Nevertheless, it is often used to decrease the variance of parameter estimates. There are two ways to choose the parameter k: cross-validation or minimizing prediction error. Hoerl and Kennard (1970) proposed the Ridge regression, and explanations of it may be found in several other texts as well (Hastie et al., 2001; Izenman, 2008). The standard Ridge regression comes in several forms. The Ridge penalized log-likelihood function can be written as the penalized least square constrained on the Ridge penalty as

$$\widehat{\boldsymbol{\beta}}_R = \arg\min_{\widehat{\boldsymbol{\beta}}} \ (\mathbf{y} - \mathbf{X})^\top (\mathbf{y} - \mathbf{X}) + k\boldsymbol{\beta}^\top \boldsymbol{\beta}/2, \tag{1.2}$$

According to Hoerl and Kennard (1970), one can easily show that equation (1.2) leads to

$$\widehat{\boldsymbol{\beta}}_R = \left(\boldsymbol{X}^T \boldsymbol{X} + k\boldsymbol{I}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}.$$

The choice of k is the biggest challenge for the Ridge regression parameter since it is crucial for controlling the bias of the regression toward the dependent variable's mean (Fayose and Ayinde, 2019). A new estimator was presented by Liu (2003) combining the strengths of the Stein estimator (Stein, 1956) with the standard Ridge regression estimator of Hoerl and Kennard (1970). The Liu-type (LT) penalized log-likelihood function can be written as the penalized least square constrained on the LT penalty as

$$\widehat{\boldsymbol{\beta}}_{LT} = \arg\min_{\beta} \ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \left[(-\frac{d}{k^{1/2}})\widehat{\boldsymbol{\beta}} - k^{1/2}\boldsymbol{\beta}\right]^\top \left[(-\frac{d}{k^{1/2}})\widehat{\boldsymbol{\beta}} - k^{1/2}\boldsymbol{\beta}\right],$$

where $k > 0$, $-\infty < d < \infty$ and $\widehat{\boldsymbol{\beta}}$ can be any estimator of $\boldsymbol{\beta}$. According to Liu

(2003), it is easy to show that the LT estimator is given by

$$\widehat{\boldsymbol{\beta}}_{LT} = \left(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}\right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{y} - d\,\widehat{\boldsymbol{\beta}}\right).$$

1.4 Mixture of Regression Models

In the regression analysis, when there are many heterogeneous groups in the population, the mixture of regression models is one of the most common techniques to incorporate the information of covariates into learning about population heterogeneity. De Veaux (1989) developed the technology of models relying on mixtures of linear regression models and, especially, to draw out the relevance of the EM algorithm to the associated maximum likelihood equations. Faria and Soromenho (2010) proposed comparing the EM algorithm, the classification EM algorithm, and the stochastic EM algorithm to estimate the coefficients of a mixture of linear regression models by maximum likelihood estimation.

Mixture of regression models is an approach to seek the heterogeneity in the response of the regression. The approach first appeared as switching regression models in economics literature (Quandt and Ramsey, 1978). It was later developed and applied in statistics and marketing to comprehend market segmentation and other facets of consumer behavior (Bai et al., 2012; Bartolucci and Scaccia, 2005). Because of the model's simplicity and efficiency in capturing non-linearity models, it has found a lot of applications, such as trajectory clustering (Gaffney and Smyth, 1999), phase retrieval (Balakrishnan et al., 2017), predictors of vehicle crashes (Zou et al., 2013), anti-psychotic induced weight gain (Nowrouzi et al., 2013) and the age

of onset of bipolar disorder (Manchia et al., 2010). According to recent research (Yi et al., 2014; Klusowski et al., 2017), there has been an interest in developing various efficient methods for estimating the parameters in the mixture of regression models under natural assumptions on the sampling distribution.

Similar to regression models, the multicollinearity problem is one of the most common problems in a mixture of regression models where the covariates are highly correlated. This problem results in unreliable maximum likelihood estimates for all coefficients of the mixture model (Inan and Erdogan, 2013; Liu, 2003).

In this thesis, we develop shrinkage methods to deal with the multicollinearity in both mixture of regression models and mixture of logistic regression models. These shrinkage methods include Ridge and LT estimators. Through extensive numerical studies, we show that the developed methods provide more reliable results in estimating the coefficients of the mixture models. We study the performance of these estimators only under multicollinearity because the Liu-type and Ridge techniques are shrinkage methods. These shrinkage methods are recommended when there is multicollinearity. In the absence of the multicollinearity problem, these methods are not recommended as they result in biased estimates.

This thesis has resulted in two papers. In the first paper, we developed shrinkage methods to estimate the parameters of the mixture of logistic regression models when there is multicollinearity. This research project has been submitted for publication (Ghanem et al., 2022a).

In the second paper, we developed Liu-type and Ridge shrinkage methods to estimate the parameters of the mixture of regression models. The performance of the shrinkage methods is evaluated via classification and stochastic versions of

EM algorithms. This research project has been submitted for publication (Ghanem et al., 2022b).

The remainder of the thesis is organized as follows. In Chapter 2, we present the estimation of the parameters of the linear regression model and logistic regression model using the ML, Ridge, and LT methods. Chapter 3 develops the shrinkage estimators for the mixture of linear regression models. Chapter 4 investigates the shrinkage estimation for the mixture of logistic regression models. We investigate the performance of the developed estimation methods through an extensive simulation and real data studies in Chapter 5. Chapter 6 presents the summary and future works.

Chapter 2

Regression Models

Regression model is one of the most popular methods that allow researchers in many fields to explain the relationship between variables. The variables' relationships can be linear or non-linear, positive or negative (Bluman, 2014). The variables in regression are divided into explanatory and response variables. The explanatory variables (or independent variables) are used to explain the changes in the response variable (or dependent variable).

Regression models are used in various applications to assist researchers in predicting responses based on a set of explanatory variables (Navidi, 2011). There are many variations of the regression models, such as simple linear regression, multiple linear regression, generalized linear regression, and non-parametric regression models. Regression modeling is essential in analyzing many medical studies, mainly observational studies. Regression model building especially includes aspects such as selecting an appropriate model type based on the nature of the outcome variable, selection of explanatory (independent) variables to include in a model, planning and carrying out model diagnostics, model validation, and model revision. Shmueli

(2010) discussed the difference between three conceptual modeling approaches: descriptive, predictive, and explanatory modeling. According to (Paldam, 2021), regression is also used in economics and management.

The logistic regression model is one of the most powerful statistical methods to predict the outcome for a binary response Y based on a set of p covariates $(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$. The logistic regression has found applications in various fields, including medical research and natural sciences. Note that the objective of logistic regression analysis is similar to any other regression model. The goal is to find the best-fitting model explaining the relationship between a set of covariates and a binary response variable. Logistic regression is categorized in a model class called a generalized linear model. The most popular method for estimating a logistic regression's parameters is a maximum-likelihood estimation (MLE). Unlike linear least squares, logistic regression does not have a closed-form expression to estimate the logistic regression's parameters. The logistic distribution is preferred for two main reasons. It is a highly flexible and easily used function from a mathematical perspective and also provides a clinically helpful interpretation. Cox and Snell (1989) discuss a few of the many distribution functions that have been suggested for use in the analysis of a dichotomous (binary) outcome variable.

Logistic regression has found many applications in medical research. Boyd et al. (1987) used logistic regression in order to predict mortality in injured patients. Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient such as age, body mass index, sex (Truett et al., 1967; Freedman, 2009). According to Palei and Das (2009), the logistic regression is also used in engineering

for predicting the probability of failure of a given process, system or product.

Maximum likelihood (ML) estimation is one of the most popular methods to estimate the coefficients of linear and logistic regression models. In the presence of multicollinearity, the ML estimates will be unreliable; ML estimators in linear regressions and logistic regressions will be inflated and may result in misleading outcomes.

This chapter will focus on linear regression and logistic regression models. We study the estimation of the model parametrers using maximum likelihood method and two shrinkage methods to deal with multicollinearity problems. These two shrinkage methods include Ridge and Liu-type (LT) estimators. This chapter is organized as follows. Section 2.1 describes the linear regression model and estimating the parameters of the model. Section 2.2 describes the logistic regression model. We also discuss the maximum likelihood estimator and how we can estimate the parameters by iterative re-weighted least square method. Sections 2.4 and 2.5 develop the Ridge and Liu-type methods in estimating the coefficients of the logistic regression.

2.1 Linear Regression

Linear regression analysis is a statistical technique that describes the relationship between a response (dependent) variable and one or more explanatory (independent) variables. Simple and multiple linear regression analyses are differentiated based on whether there are one or many explanatory variables. A simple linear regression examines how one independent variable affects one dependent variable.

In the second scenario, multiple linear regression examines the impact of various independent variables on one dependent variable.

Consider the linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\boldsymbol{X}$ is a known design matrix $n \times p$ of covariate values with rank$(\boldsymbol{X}) = p,$, $\boldsymbol{y}$ is a $n \times 1$ response vector, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)$ where ϵ_i are independent and identically normal random errors that is $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. The least-squares estimator $\widehat{\boldsymbol{\beta}}_{LS} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ and a maximum likelihood estimation are among the most popular methods used to estimate $\boldsymbol{\beta}$. The likelihood function of $\boldsymbol{\beta}$ is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \phi\left(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2\right), \tag{2.2}$$

where $\phi\left(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2\right)$ represents the pdf of normal distribution with mean $\mu_i^\top = \mathbf{x}_i^\top \boldsymbol{\beta}$ and variance σ^2. Accordingly, the log-likelihood can be obtained by

$$\ell(\boldsymbol{\beta}) = \frac{-n}{2} \log(2\pi) - \frac{n}{2} \log\left(\sigma^2\right) - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{2.3}$$

By taking the first derivative from (2.3) with respect to $\boldsymbol{\beta}$, one obtains

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Under the normality assumption of the error terms, the maximum likelihood estimation of the regression coefficient will be the same as $\widehat{\boldsymbol{\beta}}_{LS}$.

Multicollinearity happens when linear dependencies exist between the independent variables, and in this case, $\widehat{\boldsymbol{\beta}}_{LS}$ performs very poorly. If the main concern is the estimation of the regression coefficients, collinearity can severely affect least squares estimators. The variances of the least squares coefficient estimators are substantial; they may be far from the actual values, although the least squares coefficient estimators are still unbiased.

Hoerl and Kennard (1970) proposed Ridge estimator of linear regression $\widehat{\boldsymbol{\beta}}_R = \left(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ to solve the problem of collinearity and it became the most common method to cope with the weakness of least square estimator. When $\boldsymbol{X}^\top \boldsymbol{X}$ is ill-conditioned, it signifies that two or more regressors are almost linearly dependent on one other. In this situation the results of this element $\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$ become large and the variances of the estimates will increase; for this reason, $\widehat{\boldsymbol{\beta}}_{LS}$ becomes unstable. The condition number is used to measure the collinearity, and it is defined by

$$\kappa = \left(\frac{\lambda_{max}}{\lambda_{min}}\right)^{1/2}, \tag{2.4}$$

where λ_{max} and λ_{min} denotes the maximum and the minimum eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$.

A high condition number implies that $\boldsymbol{X}^\top \boldsymbol{X}$ is ill-conditioned. By adding a small constant to the diagonal of $\boldsymbol{X}^\top \boldsymbol{X}$, the Ridge estimator tries to deal with the problem of multicollinearity in order to improve its condition number. It is easy to see that the condition number of $\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_{pp}$ is a decreasing function of k. In application, the shrinkage parameter k in Ridge regression is typically relatively small. Moreover, a shrinkage parameter k should be high if we want to limit the

condition number of $\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_{pp}$ to a small level. As a result, when $\boldsymbol{X}^\top \boldsymbol{X}$ is very ill-conditioned, a small k may not be able to deal with the multicollinearity problem.

In general, the multicollinearity is not severe when the condition number, κ, is less than 10. When κ becomes greater than 100, it refers to severe multicollinearity, and the condition number between $\kappa \in$ (30,100) refers to moderate to strong multicollinearity (Belsley et al., 1980). In the case of severe multicollinearity, we have to choose large k to decrease multicollinearity. However, large values of k provide more bias to the Ridge estimator; thus, the Ridge estimator cannot completely solve the ill-conditioned design matrix.

As indicated above, Ridge estimation method uses a small values of k to deal with the multicollinearity problem. When multicollinearity is severe the small value of k will not be enough to handle the problem. On the other side, large k will dramatically add biases to the estimates. It's worth noting that Ridge regression can be obtained by adding equation $0 = k^{1/2}\boldsymbol{\beta} + \boldsymbol{\epsilon}^\top$ to the original equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and then using the least-squares estimator. The distance between $k^{1/2}\boldsymbol{\beta}$ and 0 increases as k grows. As a result, adding $0 = k^{1/2}\boldsymbol{\beta} + \boldsymbol{\epsilon}^\top$ to the original equation causes the Ridge regression to be more biased, and consequently choosing a small k is preferable.

To solve this problem, (Liu, 2003) proposed substituting $\left(-d/k^{1/2}\right)\widehat{\boldsymbol{\beta}}$ for the left side of the equation $0 = k^{1/2}\boldsymbol{\beta} + \epsilon^\top$, where $\widehat{\boldsymbol{\beta}}$ might be any estimator of $\boldsymbol{\beta}$. In the new equation $\left(-d/k^{1/2}\right)\widehat{\boldsymbol{\beta}} = k^{1/2}\boldsymbol{\beta} + \epsilon^\top$, we can use a big k since another value, d can be adjusted to make the equation fit. We get the new estimator

$\widehat{\boldsymbol{\beta}}_{LT} = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}(\boldsymbol{X}^\top \boldsymbol{y} - d\,\widehat{\boldsymbol{\beta}})$ by augmenting the new equation to $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$ and then applying the least-squares approach. This estimator is called Liu-type (LT) estimator and it is given by

$$\widehat{\boldsymbol{\beta}}_{LT} = \left(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{X} - d\boldsymbol{I}_{pp}\right) \widehat{\boldsymbol{\beta}}, \tag{2.5}$$

where $k > 0$, $-\infty < d < \infty$ and $\widehat{\boldsymbol{\beta}}$ can be the ML estimator or Ridge estimator.

The tuning parameter k can be used exclusively to regulate the condition number of $\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}$ in the estimator of $\widehat{\boldsymbol{\beta}}_{LT}$. After reducing the condition number of $\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}$ to the required amount, some bias is inevitable; thus, the second parameter d is used to improve the fit and statistical property.

When we use $\widehat{\boldsymbol{\beta}}_{LS}$ in the LT penalty (2.5), the LT estimator is given by

$$\widehat{\boldsymbol{\beta}}_{LT} = \left(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{X} - d\boldsymbol{I}_{pp}\right) \widehat{\boldsymbol{\beta}}_{LS}, \tag{2.6}$$

where $k > 0$, $-\infty < d < \infty$ and $\widehat{\boldsymbol{\beta}}_{LS}$ is the maximum likelihood estimator. There is no clear rule for choosing the tuning parameters (k,d), but according to Liu (2003), the first parameter k can be calculated in a different way, such as $p\widehat{\sigma}^2_{ML}/\widehat{\boldsymbol{\beta}}^\top_{ML}\widehat{\boldsymbol{\beta}}_{ML}$ or $\lambda_1 - 100\lambda_p/99$, where p is the number of covariates in the regression, $\widehat{\sigma}^2_{LS}$ is a mean square error of maximum likelihood estimator and λ_1, λ_p denotes the maximum and the minimum eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$. The second tuning parameter d is given by

$$\widehat{d} = \frac{\sum_{i=1}^{p} \left(\left(\widehat{\sigma}^2_{ML} - \widehat{k}\widehat{\alpha}^2_{ML,i}\right) / \left(\lambda_i + \widehat{k}\right)^2 \right)}{\sum_{i=1}^{p} \left(\left(\lambda_i \widehat{\alpha}^2_{ML,i} - \widehat{\sigma}^2_{ML}\right) / \lambda_i \left(\lambda_i + \widehat{k}\right)^2 \right)}, \tag{2.7}$$

where $\widehat{\alpha}_{ML} = \Lambda^{-1}\boldsymbol{Z}^\top \boldsymbol{y}$, such that $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{Q}$ and Q is the orthogonal matrix whose columns constitute the eigenvectors of $\boldsymbol{X}^\top \boldsymbol{X}$ and $\Lambda = \boldsymbol{Z}^\top \boldsymbol{Z} = \text{diag}\,(\lambda_1, \cdots, \lambda_p)$,where $\lambda_1 \geq \cdots \geq \lambda_p > 0$ are the ordered eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$.

When we use $\widehat{\boldsymbol{\beta}}_R$ in the LT penalty (2.5), the LT estimator is given by

$$\widehat{\boldsymbol{\beta}}_{LT} = \left(\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1} \left(\boldsymbol{X}^\top \boldsymbol{X} - d\boldsymbol{I}_{pp}\right) \widehat{\boldsymbol{\beta}}_R, \tag{2.8}$$

where $k > 0$, $-\infty < d < \infty$ and $\widehat{\boldsymbol{\beta}}_R$ is the Ridge estimator. Readers are referred to Liu (2003) for more details about the LT estimator of the coefficients of the linear regression model.

2.2 Logistic Regression

The logistic regression model is a popular method utilized in various fields, including medical research and natural sciences. The logistic regression model uses the information from a collection of explanatory independent variables to explain the binary response variable which has only two possible outcomes, for instance, success or failure of an experiment, presence or absence of disease.

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the vector of binary responses from a random sample of size n. Let $\boldsymbol{X}$ represent non-random $(n \times p)$ design matrix of p explanatory variables $(\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ with $\text{rank}(\boldsymbol{X}) = p < n$. The logistic regression model is then given by

$$\mathbb{P}(y_i = 1|\mathbf{X}) = p(\mathbf{x}_i; \boldsymbol{\beta}) = 1/\left(1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})\right), \tag{2.9}$$

where $\boldsymbol{\beta}$ represents the vector of the unknown coefficients of the logistic regression.

Instead of utilizing $\mathbf{y}$ as the dependent variable in the logistic regression function, there is an alternative function called logit, which is defined as the natural logarithmic of the odds and it is given by $\ln\left(\frac{p(\mathbf{x}_i;\boldsymbol{\beta})}{1-p(\mathbf{x}_i;\boldsymbol{\beta})}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}$. The probability determines the value of the logit function. Note that the logistic regression function takes a value between $[0, 1]$, while logit function can be any real number between $[-\infty, \infty]$.

Generalized linear models (GLMs) are a natural generalization of classical linear models that allow a population's mean to depend on a linear predictor through a (possibly nonlinear) link function. This allows the response probability distribution to be any member of the exponential family of distributions. Logistic regression is one special case of generalized linear models. The random component of the model is given by $Y_i \overset{iid}{\sim} B\left(1, p(\mathbf{x}_i;\boldsymbol{\beta})\right)$. Likelihood method is a popular method to estimate the parameter of the logit model. The parameters $\boldsymbol{\beta}_j$, $j = 1, \cdots, p$ are interpreted as the log odds ratio of $(y_i = 1)$ when $\boldsymbol{x}_j$ changes by one unit.

2.3 Maximum Likelihhood Estimator

Maximum likelihood (ML) estimation is one of the most popular methods for estimating the parameters of linear and logistic regression models. The ML method does not impose any restrictions on the independent variables. Let Y_i be the binary response variable associated to the i-th subject, where $Y_i \overset{iid}{\sim} B\left(1, p(\mathbf{x}_i;\boldsymbol{\beta})\right)$; $i = 1, \cdots, n$, where $p(\mathbf{x}_i;\boldsymbol{\beta})$ is given by (2.9). The likelihood function of $\boldsymbol{\beta}$ is then

given by

$$L\left(\boldsymbol{\beta}|D\right)=\prod_{i=1}^{n}\left\{\left(\frac{1}{1+e^{-\mathbf{x}_i^{\top}\boldsymbol{\beta}}}\right)^{y_i}\left(1-\frac{1}{1+e^{-\mathbf{x}_i^{\top}\boldsymbol{\beta}}}\right)^{1-y_i}\right\}.$$

Accordingly, the log-likelihood can be obtained by

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^{n}\left\{\mathbf{x}_i^{\top}\boldsymbol{\beta}y_i-\log\left(1+\exp(-\mathbf{x}_i^{\top}\boldsymbol{\beta})\right)\right\}. \tag{2.10}$$

It is known that there is no closed-form solution to the maximum likelihood estimate of $\boldsymbol{\beta}$ in the logistic regression model. Thus, Newton-Raphson (NR) technique is typically used to estimate the coefficients of the logistic regression (2.9). The NR algorithm iteratively estimate $\boldsymbol{\beta}$ as follows:

$$\boldsymbol{\beta}^{(t+1)}=\boldsymbol{\beta}^{(t)}-H^{-1}(\ell)\left(\boldsymbol{\beta}^{(t)}\right).\nabla_{\boldsymbol{\beta}}\ell\left(\boldsymbol{\beta}^{(t)}\right), \tag{2.11}$$

where $\boldsymbol{\beta}^{(t)}$ is the estimate updated from iteration t. Also $\nabla_{\boldsymbol{\beta}}\ell\left(\boldsymbol{\beta}^{(t)}\right)$ and $H^{-1}(\ell)\left(\boldsymbol{\beta}^{(t)}\right)$ represent respectively the gradient and hessian matrix evaluated at $\boldsymbol{\beta}^{(t)}$. By taking the first derivative from (2.10) with respect to β_l, the gradient is obtained by

$$\begin{aligned}\frac{\partial\ell\left(\boldsymbol{\beta}\right)}{\partial\beta_l}&=\sum_{i=1}^{n}\left\{-\left(\frac{e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}}\right)x_{il}+\left(y_ix_{il}\right)\right\}\\&=\sum_{i=1}^{n}\left(y_i-p_i\right)x_{il},\end{aligned} \tag{2.12}$$

where

$$p_i=\frac{e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}}. \tag{2.13}$$

We can write the gradient (2.12) in a matrix form as

$$\nabla_{\boldsymbol{\beta}} \ell\left(\boldsymbol{\beta}\right) = \boldsymbol{X}^{\top}\left(\boldsymbol{y} - \boldsymbol{p}\right). \tag{2.14}$$

Once the gradient is obtained in (2.14), we need to focus on the second derivative of (2.10) with respect to $\boldsymbol{\beta}$ and calculate the Hessian matrix. Accordingly, the (k, l) entry of the Hessian matrix is calculated by

$$\begin{aligned}\frac{\partial^2 \ell\left(\boldsymbol{\beta}\right)}{\partial \beta_k \partial \beta_l} &= \sum_{i=1}^{n} \left\{ -x_{ik} x_{il} \left(\frac{1}{1 + e^{-\boldsymbol{x}_i^{\top}\boldsymbol{\beta}}} \right)^2 e^{-\boldsymbol{x}_i^{\top}\boldsymbol{\beta}} \right\} \\ &= -\sum_{i=1}^{n} x_{il} x_{ik} p_i \left(1 - p_i\right) \end{aligned} \tag{2.15}$$

where p_i is obtain from (2.13). The Hessian matrix can be written as the inner-product of the weighted matrix by

$$H\left(\ell\right) = -\boldsymbol{X}^{\top}\boldsymbol{D}\boldsymbol{X}, \tag{2.16}$$

where $\boldsymbol{D}$ is a diagonal matrix with $\boldsymbol{D}_{ii} = p_i\left(1 - p_i\right); i = 1, \cdots, n$. Now from the gradient function (2.14) and the hessian matrix (2.16), one can use the NR method and obtain the ML estimate of the coefficients of the logistic regression by

$$H(\ell)\left(\boldsymbol{\beta}^{(t)}\right)\left(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)}\right) = -\nabla_{\boldsymbol{\beta}} \ell\left(\boldsymbol{\beta}^{(t)}\right).$$

2.3.1 Iterative Re-weighted Least Squares Method

The NR can be reformulated as an iterative re-weighted least squares method. we can re-write the gradient (2.14) and hessian (2.16) in matrix form as follows

$$\frac{\partial \ell}{\partial \beta} = \boldsymbol{X}^{\top}\left(\boldsymbol{y} - g^{-1}\left(\boldsymbol{x};\boldsymbol{\beta}\right)\right),$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^{\top}} = -\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X},$$

where $g^{-1}\left(\boldsymbol{x};\boldsymbol{\beta}\right) = \left[g^{-1}\left(\boldsymbol{x}_1;\boldsymbol{\beta}\right), ..., g^{-1}\left(\boldsymbol{x}_n;\boldsymbol{\beta}\right)\right]^{\top}$, and $g^{-1}(\mathbf{x}_i;\boldsymbol{\beta})$ is given by

$$g^{-1}(\mathbf{x}_i;\boldsymbol{\beta}) = 1/\left(1+\exp(-\mathbf{x}_i^{\top}\boldsymbol{\beta})\right),$$

and $\boldsymbol{W}$ diagonal matrix with $\boldsymbol{W}_{ii} = \dfrac{e^{-\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}}}{\left(1+e^{-\boldsymbol{x}_i^{\top}\widehat{\boldsymbol{\beta}}}\right)^2}, i = 1, ..., n$. Hence the equation (2.11) can be updated as follows

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{new} &= \widehat{\boldsymbol{\beta}}^{old} + \left(\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\left[\boldsymbol{y} - g^{-1}\left(\boldsymbol{x},\widehat{\boldsymbol{\beta}}^{old}\right)\right] \\
&= \left(\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{W}\left\{\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{old} + \boldsymbol{W}^{-1}\left[\boldsymbol{y} - g^{-1}\left(\boldsymbol{x},\widehat{\boldsymbol{\beta}}^{old}\right)\right]\right\} \\
&= \left(\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{Z}, \qquad (2.17)
\end{aligned}$$

where $\boldsymbol{Z} = \left\{\boldsymbol{X}\widehat{\boldsymbol{\beta}}^{old} + \boldsymbol{W}^{-1}\left[\boldsymbol{y} - g^{-1}\left(\boldsymbol{x};\widehat{\boldsymbol{\beta}}^{old}\right)\right]\right\}$. The Newton-Raphson Algorithm update is thus the solution to the following weighted least squares problem.

$$\widehat{\boldsymbol{\beta}}^{new} = \arg\min_{\beta}\left(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta}\right)^{\top}\boldsymbol{W}\left(\boldsymbol{Z} - \boldsymbol{X}\boldsymbol{\beta}\right). \qquad (2.18)$$

Effectively, at each iteration, the adjusted response $\boldsymbol{Z}$ is regressed on the covariates including $\boldsymbol{X}$. Comparing equations (2.17) and (2.18), we can view (2.17) as the $\widehat{\boldsymbol{\beta}}$ estimator of weighted regression $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\beta}$.

Based on regularity conditions of (Ngunyi et al., 2014; Beer, 2001; Rashid and Shifa, 2009), as $n \to \infty$, the MLE estimator will be asymptotic consistent estimator for $\boldsymbol{\beta}$ and $\widehat{\beta}_{jn}$ is consistent for β_j and $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\right)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $[I(\boldsymbol{\beta})^{-1}]$ such that $\sqrt{n}\left(\widehat{\beta}_{jn} - \beta\right) \to N\left(0, [I(\boldsymbol{\beta})^{-1}]_{jj}\right)$, where $\widehat{\beta}_{jn}$ represents the $j-$th element of $\widehat{\boldsymbol{\beta}}_n$, $I(\boldsymbol{\beta})$ is Fisher's information matrix and $[I(\boldsymbol{\beta})^{-1}]_{jj}$ is the $j-$th element of the inverse of Fisher's information matrix. For more details, see (Ngunyi et al., 2014).

2.4 Ridge Estimator of Logistic Regression Parameters

Hoerl and Kennard (1970) proposed the method of Ridge regression, as the result of derived from a restricted maximum likelihood method (REML).

In this section, we derive the Ridge estimation of logistic regression coefficients. Ridge estimator of the logistic regression is found by the maximization of the Ridge penalized likelihood (Duffy and Santner, 1989). The Ridge penalized log likelihood is given by

$$\ell^R(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - k\|\boldsymbol{\beta}\|_2^2, \tag{2.19}$$

where $\ell(\boldsymbol{\beta})$ is the unrestricted log-likelihood function of logistic regression and $\|\boldsymbol{\beta}\| = \left(\sum_j \beta_j^2\right)^{1/2}$. From equation (2.19), the Ridge penalized log-likelihood can

be written by

$$\ell^R(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\boldsymbol{x}_i^\top \boldsymbol{\beta} y_i - \log\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_j}\right)\right) - \frac{1}{2} k \boldsymbol{\beta}\boldsymbol{\beta}^\top.$$

By taking the first and second derivatives from the log-likelihood function (2.19), the gradient and Hessian matrix are given by

$$\begin{aligned} \frac{\partial \ell^R(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - k\boldsymbol{\beta} \\ &= \boldsymbol{X}^\top \left(\boldsymbol{y} - g^{-1}(\boldsymbol{x}; \boldsymbol{\beta})\right) - k\boldsymbol{\beta}, \end{aligned} \tag{2.20}$$

$$\begin{aligned} \frac{\partial^2 \ell^R(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} - k\boldsymbol{I} \\ &= -\left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + k\boldsymbol{I}\right) = -\boldsymbol{V}. \end{aligned} \tag{2.21}$$

Let $\boldsymbol{V} = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + k\boldsymbol{I}_{\boldsymbol{pp}}$. From eqns (2.20) and (2.21), the estimation of $\boldsymbol{\beta}$ based on the penalized likelihood can be obtained by the following NR iteration:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{new} &= \widehat{\boldsymbol{\beta}}^{old} + \boldsymbol{V}^{-1} \left\{ \boldsymbol{X}^\top \left(\boldsymbol{y} - g^{-1}\left(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}^{old}\right)\right) - k\widehat{\boldsymbol{\beta}}^{old} \right\} \\ &= \boldsymbol{V}^{-1} \boldsymbol{V} \widehat{\boldsymbol{\beta}}^{old} - k\boldsymbol{V}^{-1} \widehat{\boldsymbol{\beta}}^{old} + \boldsymbol{V}^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{W}^{-1} \left\{ \left(\boldsymbol{y} - g^{-1}\left(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}^{old}\right)\right)\right\} \\ &= \boldsymbol{V}^{-1} \boldsymbol{X}^\top \boldsymbol{W} \left\{ \boldsymbol{X} \widehat{\boldsymbol{\beta}}^{old} + \boldsymbol{W}^{-1} \left[\boldsymbol{y} - g^{-1}\left(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}^{old}\right)\right]\right\} \\ &= \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Z}, \end{aligned} \tag{2.22}$$

where

$$W_{ii} = \frac{e^{-\boldsymbol{x_i}^\top \widehat{\boldsymbol{\beta}}}}{\left(1 + e^{-\boldsymbol{x_i}^\top \widehat{\boldsymbol{\beta}}}\right)^2}$$

$$\boldsymbol{Z} = \left\{ \boldsymbol{X}\, \widehat{\boldsymbol{\beta}}^{old} + \boldsymbol{W}^{-1} \left[\boldsymbol{y} - g^{-1}\left(\boldsymbol{x}; \widehat{\boldsymbol{\beta}}^{old} \right) \right] \right\},$$

and $\boldsymbol{I}_{pp}$ is the identity matrix with size $(p \times p)$. When there is multicollinearity, the mean square error of the Ridge logistic estimator is smaller than the ML estimator (Schaefer et al., 1984). From equation (2.22), the Ridge logistic estimator can be written by

$$\widehat{\boldsymbol{\beta}}_R = \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + k \boldsymbol{I}_{pp}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\, \widehat{\boldsymbol{\beta}}_{ML}, \tag{2.23}$$

where $\widehat{\boldsymbol{\beta}}_{ML} = \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Z}$ is the ML estimate of $\boldsymbol{\beta}$.

Like the Ridge regression estimator, a high condition number implies that $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ is ill-conditioned. When the condition number of $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ is high, the effect of collinearity on the least squares estimator is most serious. The Ridge estimator could deal with this problem by adding a small constant to the diagonal of $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ to improve its condition number. The condition number is given in equation (2.4), where λ_{max} and λ_{min} denotes the maximum and the minimum eigenvalues of $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$.

2.5 Liu-type Estimator for Logistic Regression Parameters

The Liu-type (LT) logistic estimator was presented to address the problem of extreme multicollinearity, with the expectation of a smaller MSE than the Ridge

logistic regression. The LT estimator of coefficients of logistic is defined as

$$\widehat{\boldsymbol{\beta}}_{LT} = \left(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1}\left(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X} - d\boldsymbol{I}_{pp}\right)\widehat{\boldsymbol{\beta}}, \tag{2.24}$$

where $k > 0$, $-\infty < d < \infty$ and $\widehat{\boldsymbol{\beta}}$ can be the maximum likelihood (ML) estimator or Ridge estimator. The LT estimation method requires two tuning parameters. The first parameter is k, which is designed to control the value of condition number of $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X} + k\boldsymbol{I}_{pp}$. The inevitable bias caused by k can be adjusted with a second parameter d, the so-called bias correction parameter after the condition number of $\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X}+k\boldsymbol{I}_{pp}$ has been decreased to the desired amount. These tuning parameters will enable the LT estimator to deal with the problem of severe multicollinearity. Hence, the LT estimator will yield a smaller MSE than the Ridge method in estimating the parameters of the logistic regression. When we use $\widehat{\boldsymbol{\beta}}_R$ in the LT penalty (2.24), the LT logistic estimator is obtained by

$$\widehat{\boldsymbol{\beta}}_{LT} = \left(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X} + k\boldsymbol{I}_{pp}\right)^{-1}\left(\boldsymbol{X}^\top \boldsymbol{W}\boldsymbol{X} - d\boldsymbol{I}_{pp}\right)\widehat{\boldsymbol{\beta}}_R, \tag{2.25}$$

where $k > 0$, $-\infty < d < \infty$ and $\widehat{\boldsymbol{\beta}}_R$ is the Ridge logistic estimator given by (2.23). There is no clear rule for choosing k, but for logistic regression models, there are many choices of Ridge parameter such as $1/\widehat{\boldsymbol{\beta}}_{ML}^\top\widehat{\boldsymbol{\beta}}_{ML}, p/\widehat{\boldsymbol{\beta}}_{ML}^\top\widehat{\boldsymbol{\beta}}_{ML}, (p+1)/\widehat{\boldsymbol{\beta}}_{ML}^\top\widehat{\boldsymbol{\beta}}_{ML}$ (Schaefer et al., 1984; Smith et al., 1991), where p is the number of covariates in the regression. From (2.25), the MSE of the LT logistic estimator is obtained by

$$\begin{aligned}\text{MSE}\left(\widehat{\boldsymbol{\beta}}_{LT}\right) = tr\Big[&\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+k\boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}-d\boldsymbol{I}_p\right)\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}\right)^{-1}\\ &\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}-d\boldsymbol{I}_p\right)\left(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}+k\boldsymbol{I}_p\right)^{-1}\Big]\\ &+\Big\|\left(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}+k\boldsymbol{I}_p\right)^{-1}\left(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}-d\boldsymbol{I}_p\right)\left(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\right)^{-1}\\ &\boldsymbol{X}^\top\boldsymbol{W}p\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)-\boldsymbol{\beta}\Big\|_2^2.\end{aligned} \tag{2.26}$$

For a fixed k, the MSE($\widehat{\boldsymbol{\beta}}_{LT}$) is a quadratic function of d according to Inan and Erdogan (2013). Therefore, it is straightforward to find the optimum d value that minimizes the MSE given by (2.26). If the optimum d value is utilized as a proposed estimator, the MSE($\widehat{\boldsymbol{\beta}}_{LT}$) is always less than or equal to the Ridge estimator's MSE. From (2.26), we can see MSE($\widehat{\boldsymbol{\beta}}_{LT}$) depends on two unknown parameters and thus we can not calculate the MSE in this case. Inan and Erdogan (2013) proposed a solution in order to calculate the MSE($\widehat{\boldsymbol{\beta}}_{LT}$) for any practical study. This solution depends on replacing the unknown parameter $\boldsymbol{\beta}$ with its estimate $\widehat{\boldsymbol{\beta}}_{\boldsymbol{ML}}$ and $p\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)$ with $p(\boldsymbol{x}_i;\widehat{\boldsymbol{\beta}})$ in (2.26).

Chapter 3

Shrinkage Estimators for Mixture of Linear Regressions

In this chapter, we focus on a finite mixture of regression models. Multicollinearity significantly impacts the Maximum likelihood (ML) estimate of the mixture of regression models, just as it does on the regression model. We develop two shrinkage approaches through an unsupervised learning approach to estimate the model coefficients even in multicollinearity issues (Ghanem et al., 2022b). These approaches include the Liu-type (LT) and Ridge shrinkage estimation methods. The performance of the developed methods is evaluated via classification and stochastic versions of EM algorithms. We show that the LT estimators outperform their Ridge and ML counterparts in estimating the coefficients of the mixture of regression models through various numerical studies in Chapter 5.

This chapter is organized as follows. Section 3.1 presents an introduction to the mixture of regression models. Section 3.2 describes the statistical part of mixture of the regression models. Sections 3.3, 3.4 and 3.5 describe the ML, Ridge and LT methods in estimating the parameters of the mixture of regression models.

3.1 Introduction

The finite mixture models have been used to study various random occurrences. A mixture model is a probabilistic model that identifies the sub-populations within a larger population. A mixture model formally proposes a mixture of distributions to model the heterogeneity of observations in the population. Finite mixture models have found applications in various disciplines, including genetics, economics, and medicine (Lindsay, 1995; Böhning, 1999).

The mixture of linear regressions is one of the popular methods in mixture modeling. The mixture of regressions is vital when no information matches the observations to the component regressions. Quandt and Ramsey (1978) have proposed a general form for a mixture of the linear regression models, namely switching regression. In order to estimate the parameters, the method relied on the definition of the moment-generation function.

The expectation-maximization (EM) method was developed to fit the two regression problems by De Veaux (1989). Jones and McLachlan (1992) used the EM method in order to fit these two regression models and applied mixture of regressions in data analysis. The two-component mixture of single variable linear regression has been fitted using the EM algorithm (Turner, 2000). Hawkins et al. (2001) considered the determining the number of components in the mixture of linear regression models utilizing the likelihood equation. Also, the asymptotic theory for maximum likelihood estimator has been investigated for mixture regression models by Zhu and Zhang (2004).

In this chapter, we investigate the approaches to fit a mixture of linear regressions

by the concept of the maximum likelihood method. We discuss three approaches for maximization to achieve the maximum likelihood estimators. These approaches include the expectation-maximization algorithm (EM) (Dempster et al., 1977), the classification EM algorithm (CEM) (Celeux and Govaert, 1992) and the stochastic EM algorithm (SEM) (Celeux, 1985). Ganesalingam (1989) proposed numerical studies to compare the EM and CEM techniques using the Gaussian mixture in practical cases. Celeux and Govaert (1993) extended this analysis to clarify the impact of sample sizes and the algorithms' dependence on their starting values. Celeux and Govaert (1993) proposed an extension of the comparisons to Bernoulli models in the case of binary data. There are also some comparisons of EM and SEM approaches in various distributions. Celeux et al. (1996) used Monte Carlo numerical simulations and real data to compare these techniques. Dias and Wedel (2004) examined EM and SEM techniques for estimating Gaussian mixture model parameters.

3.2 Statistical Method

Regression model $y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ is one of the most popular statistical methods to study the relationship between response variable and the independent variables in the design matrix. Let $\mathbf{x}_i^\top = (x_{i,1}, \ldots, x_{i,p})$ be the vector of p independent variables for the i-th subject in a random sample of size n. Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the vector of responses from a sample of size n. Let $\mathbf{X}$ denote $(n \times p)$ design matrix of p independent variables of $\text{rank}(\mathbf{X}) = p < n$.

The mixture of regression models is a generalization of the regression model

when the observed data come from M components.While the number of components M is assumed to be known throughout this thesis, the problem is treated as an unsupervised learning approach when the component membership of observations are unknown and should be estimated. The mixture of regression models is given by

$$y_i = \begin{cases} \mathbf{x}_i^\top \boldsymbol{\beta}_1 + \epsilon_{i1} & \text{with probability } \pi_1, \\ \mathbf{x}_i^\top \boldsymbol{\beta}_2 + \epsilon_{i2} & \text{with probability } \pi_2, \\ \vdots & \vdots \\ \mathbf{x}_i^\top \boldsymbol{\beta}_M + \epsilon_{iM} & \text{with probability } \pi_M, \end{cases} \tag{3.1}$$

where ϵ_{ij} be a random variables with $\epsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma_j^2\right), i = 1, ..., n$ and $\boldsymbol{\beta}_j = (\beta_{j,1}, \cdots, \beta_{j,p})$ represent the coefficients of p predictors in the j-th component regression for $j = 1, \cdots, M$ and $\pi = (\pi_1, \ldots, \pi_M)$ denote the vector of the mixing proportions with $\pi_j > 0$ and $\sum_{j=1}^{M} \pi_j = 1$. Let $\boldsymbol{\theta}_j = \left(\boldsymbol{\beta}_j, \sigma_j^2\right)$ represent the parameters of the j-th component. Thus, we donate the vector of all unknown parameters of mixture (3.1) with $\boldsymbol{\Psi} = (\pi, \boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_M)$.

From regression model (3.1), the log-likelihood function of $\boldsymbol{\Psi}$ can be written as

$$\ell\left(\boldsymbol{\Psi}\right) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{M} \pi_j \phi_j \left(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right) \right), \tag{3.2}$$

where $\phi_j \left(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right)$ represents the univariate normal distribution with mean $\mu_j = \mathbf{x}_i^\top \boldsymbol{\beta}_j$ and variance σ_j^2. To get the ML estimate of $\boldsymbol{\Psi}$, we must maximize the log-likelihood function (3.2). The gradient of (3.2) is not tractable with respect to component parameters $\boldsymbol{\theta}_j, j = 1, \cdots, M$. We consider $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ to be incomplete data and use the expectation-maximization (EM) approach from (3.2) to

determine $\widehat{\boldsymbol{\Psi}}_{ML}$. Suppose $\{(\mathbf{x}_i, y_i, \mathbf{Z}_i), i = 1, \ldots, n\}$ denote the complete data where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ is the latent variable repressing the component membership of the i-th subject with

$$Z_{ij} = \begin{cases} 1 & \text{if the } i\text{-th subject comes from the } j\text{-th component,} \\ 0 & o.w., \end{cases}$$

where $\mathbf{Z}_i \overset{iid}{\sim} \text{Multi}(1, \pi_1, \ldots, \pi_M)$. The conditional distribution of $\mathbf{Z}_i|y_i$ is calculated using the marginal distribution of the latent variables and it is given by

$$f(\boldsymbol{z}_i|y_i) = \prod_{j=1}^{M} \left\{ \frac{\pi_j \phi_j\left(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right)}{\sum_{j=1}^{M} \pi_j \phi_j\left(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right)} \right\}^{z_{ji}}. \tag{3.3}$$

From above, it is easy to show $\mathbf{Z}_i|y_i \overset{iid}{\sim} \text{Multi}(1, \tau_{i1}(\boldsymbol{\Psi}), \ldots, \tau_{iM}(\boldsymbol{\Psi}))$ where

$$\tau_{ij}(\boldsymbol{\Psi}) = \frac{\pi_j \phi_j\left(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right)}{\sum_{j=1}^{M} \pi_j \phi_j\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right)}. \tag{3.4}$$

Thus, the complete log-likelihood of $\boldsymbol{\Psi}$ is given by

$$\ell_c(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \log(\pi_j) + \sum_{i=1}^{n} \sum_{j=1}^{M} z_{ij} \log \phi_j\left(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right). \tag{3.5}$$

3.3 ML Estimation Method

The expectation maximization (EM) algorithm is a well-known technique to find the maximum likelihood estimate of mixture model parameters. The EM algorithm decomposes the estimation process into iterative expectation (E) and maximization (M) steps using latent variables on top of the observed data. As an iterative method,

EM algorithm starts with an initial values. Let $\mathbf{\Psi}^{(0)} = (\pi^{(0)}, \boldsymbol{\theta}_1^{(0)}, \cdots, \boldsymbol{\theta}_M^{(0)})$ and $\boldsymbol{\Psi}^{(r)}$ represent the initial values and the estimate in the r-th iteration of the EM algorithm, respectively. We have to compute the conditional expectation of the entire log-likelihood function (3.5) in the E-step on the $(r+1)-th$ iteration. The conditional log-likelihood function $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(r)})$ replaces the latent variables by the conditional expectation of the latent variables as

$$\mathbf{Q}(\mathbf{\Psi}, \mathbf{\Psi}^{(r)}) = \mathbf{Q}_1(\pi, \mathbf{\Psi}^{(r)}) + \mathbf{Q}_2(\boldsymbol{\theta}, \mathbf{\Psi}^{(r)}),$$

where

$$\mathbf{Q}_1(\pi, \mathbf{\Psi}^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{M} \tau_{ij}(\boldsymbol{\Psi}^{(r)}) \log\left(\pi_j\right), \tag{3.6}$$

and

$$\mathbf{Q}_2(\boldsymbol{\theta}, \mathbf{\Psi}^{(r)}) = \sum_{i=1}^{n} \sum_{j=1}^{M} \tau_{ij}\left(\boldsymbol{\Psi}^{(r)}\right) \log \phi_j\left(\mathbf{x}_i^\top \boldsymbol{\beta}_j, \sigma_j^2\right), \tag{3.7}$$

where $\tau_{ij}(\boldsymbol{\Psi}^{(r)})$ is achived by (3.4). In the M-step, we must maximize $Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(r)})$ with respect to π and $\boldsymbol{\theta}$ in order to update $\boldsymbol{\Psi}^{(r+1)}$. One can update $\pi^{(r+1)}$ by maximizing $\mathbf{Q}_1(\pi, \mathbf{\Psi}^{(r)})$ subject to $\sum_{j=1}^{M} \pi_j = 1$ as follows

$$\widehat{\pi}_j^{(r+1)} = \sum_{i=1}^{n} \tau_{ij}(\mathbf{\Psi}^{(r)})/n; \qquad j = 1, \ldots, M-1. \tag{3.8}$$

The maximization of $\mathbf{Q}_2(\boldsymbol{\theta}, \mathbf{\Psi}^{(r)})$ can be reformulated as the weighted least square method as follows

$$\widehat{\boldsymbol{\beta}}_j^{(r+1)} = \arg\min_{\boldsymbol{\beta}_j} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}_j (\mathbf{y} - \mathbf{X}\beta)/n, \tag{3.9}$$

where $\mathbf{W}_j$ is $n \times n$ diagonal matrix with diagonal elements $(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{nj}(\boldsymbol{\Psi}^{(r)}))$ for all $j = 1, \ldots, M$. One can easily update $\widehat{\boldsymbol{\beta}}_j^{(r+1)}$ as the solution to (3.9) by

$$\widehat{\boldsymbol{\beta}}_j^{(r+1)} = \left(\mathbf{X}^\top \mathbf{W}_j \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W}_j \mathbf{y}, \quad j = 1, \ldots, M. \tag{3.10}$$

From the weighted least square (3.9) and following (Faria and Soromenho, 2010), we then update $\widehat{\sigma}_j^{2(r+1)}$ as follows

$$\widehat{\sigma}_j^{2(r+1)} = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{(r+1)})}{\sum_{i=1}^{n} \tau_{ij}(\boldsymbol{\Psi}^{(r)})}, \quad j = 1, \ldots, M. \tag{3.11}$$

To find $\widehat{\boldsymbol{\Psi}}_{ML}$, we iteratively alternate the E- and M- steps of the EM algorithm until the stopping criterion $|\ell(\boldsymbol{\Psi}^{(r+1)}) - \ell(\boldsymbol{\Psi}^{(r)})|$ becomes negligible.

3.3.1 Classification EM Algorithm

In the EM algorithm mentioned above, we estimate the component parameters of the mixture of regressions using information from all observations (as membership probabilities) in each iteration. Following Celeux and Govaert (1992), we will estimate $\boldsymbol{\Psi}$ iteratively using the classification version of the EM algorithm (CEM). The CEM technique includes a classification (C) step between the E- and M-steps, which updates the mixture's component parameters using the classified complete data log-likelihood function in the M-step.

The E-step here is the same as E-step of the EM algorithm. In C-step, the observations are then assigned to M mutually exclusive partitions corresponding to the M components of mixture model (3.1). Let $\mathbf{P}^{(r+1)} = (P_1^{(r+1)}, \ldots, P_M^{(r+1)})$

represent the partition in the $(r+1)$-th iteration. Each subject $(\boldsymbol{x}_i, y_i)$ is assigned to partition $P_h^{(r+1)}$ when

$$\tau_{ih}(\boldsymbol{\Psi}^{(r)}) = \arg\min_j \tau_{ij}(\boldsymbol{\Psi}^{(r)}).$$

Note that if the maximum weight isn't unique, the tie will be broken at random. When a partition becomes empty or contains only one observation, the CEM algorithm is also terminated and $\boldsymbol{\Psi}^{(r)}$ is returned.

In the M-step, we maximize conditional expectation of complete log-likelihood using the partition $P^{(r+1)}$. From (3.6) the mixing proportion is updated by

$$\widehat{\pi}_j^{(r+1)} = n_j/n, \quad j = 1, \ldots, J, \tag{3.12}$$

where n_j is the number observations allocated to partition $P_j^{(r+1)}$. Applying the weighted least square (3.9) to each partition $P_j^{(r+1)}$, we can update the parameters of the j-th component $j = 1, \ldots, M$ by

$$\widehat{\boldsymbol{\beta}}_j^{(r+1)} = \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j\right)^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{y}_j, \tag{3.13}$$

$$\widehat{\sigma}_j^{2(r+1)} = \frac{(\mathbf{y}_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}}^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y}_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}}^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\boldsymbol{\Psi}^{(r)})}, \tag{3.14}$$

where $\mathbf{X}_j$ and $\mathbf{y}_j$ represent, respectively, $(n_j \times p)$ design matrix and vector of responses corresponding to observations allocated to $P_j^{(r+1)}$. Also $\mathbf{W}_j^{(r)}$ is the diagonal weight matrix of size n_j with diagonal entries $(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{n_j,j}(\boldsymbol{\Psi}^{(r)}))$. Finally, we alternate repeatedly the E-, C- and M- steps until $|\ell(\boldsymbol{\Psi}^{(r+1)}) - \ell(\boldsymbol{\Psi}^{(r)})|$ becomes negligible.

3.3.2 Stochastic EM Algorithm

To fit a mixture of regression models, the stochastic version of the EM method (Celeux, 1985) might be used. In each iteration, the stochastic EM (SEM) method implements a stochastic version of the S-step between E- and M-steps. Despite the fact that the E- and M-steps in the SEM algorithm are the same to those in the CEM algorithm, the SEM simulates a realization of the unobserved indicator $z_i; i = 1, ..., n$ by drawing them at random from their current conditional distribution as

$$\mathbf{Z}_i^* = (\mathbf{Z}_{i1}^*, \cdots, \mathbf{Z}_{iM}^*) \overset{iid}{\sim} \text{Multi}(1, \tau_{i1}(\mathbf{\Psi}^{(r)}), \ldots, \tau_{iM}(\mathbf{\Psi}^{(r)}).$$

Then observations $(\boldsymbol{x}_i, y_i)$ is classified to partition $P_j^{(r+1)}$ if $\mathbf{Z}_{ij}^* = 1, i = 1, \cdots, n, j = 1, \cdots, M$. Using the stochastic partitions developed in S-step, we update the mixture parameters from (3.12), (3.13) and (3.14) in M-step.

From Celeux (1985) and Faria and Soromenho (2010), point-wise convergence in SEM is not guaranteed. The estimation technique resembles a Markov chain in which the maximum likelihood estimate moves around a stationary state. To do this, we alternate the E-, S-, and M-steps until the criterion $|\ell(\mathbf{\Psi}^{(r+1)}) - \ell(\mathbf{\Psi}^{(r)})|$ becomes negligible (similar to the same algorithms) or the chain exceeds a pre-specified maximum number of iterations, which is fixed for all EM, CEM, and SEM algorithms for fair comparison.

3.4 Ridge Estimation Method

The ML method is a specific tool to estimate the parameters of the mixture of regression models; however, when the covariates are linearly correlated, the ML

estimates are typically affected by multicollinearity. The Ridge estimation method is one of the most common ways of dealing with the issues of least square regression (Hoerl and Kennard, 1970). The Ridge estimate for the parameters of mixture (3.1) can be obtained as a solution to the penalized log-likelihood function given by

$$\ell^R(\boldsymbol{\Psi}) = \ell(\boldsymbol{\Psi}) - k\boldsymbol{\beta}^\top\boldsymbol{\beta}/2 \tag{3.15}$$

where $\ell(\boldsymbol{\Psi})$ is the incomplete log-likelihood from (3.2) and $k > 0$ is the Ridge parameter. In the same manner as Subsection 3.3, for each observation $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, we first introduce M dimensional latent vector $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM})$. The Ridge estimate of $\boldsymbol{\Psi}$ is then obtained using an EM technique that maximises the full Ridge log-likelihood function.

The E-step of the Ridge EM algorithm is the same as the E-step of Subsection 3.3. In the M-step, the mixing proportion are updated from (3.8). To update the coefficients of the component regressions, we require to maximize $\mathbf{Q}_2(\boldsymbol{\theta}, \boldsymbol{\Psi}^{(r)})$ subject to the Ridge penalty within each component of the mixture as

$$\mathbf{Q}_2^R(\boldsymbol{\theta}, \boldsymbol{\Psi}^{(r)}) = \mathbf{Q}_2(\boldsymbol{\theta}, \boldsymbol{\Psi}^{(r)}) - k_j\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j/2,$$

where $\mathbf{Q}_2(\boldsymbol{\theta}, \boldsymbol{\Psi}^{(r)})$ is from (3.7) and k_j is the Ridge parameter in the j-th component. Like ML method, one can write the maximization of $\mathbf{Q}_2^R(\boldsymbol{\theta}, \boldsymbol{\Psi}^{(r)})$ on the $(r+1)$-th iteration to a weighted least square subject to Ridge penalty as

$$\widehat{\boldsymbol{\beta}}_{R,j}^{(r+1)} = \arg\min_{\boldsymbol{\beta}_j} \ (\mathbf{y} - \mathbf{X})^\top\mathbf{W}_j(\mathbf{y} - \mathbf{X}) + k_j\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j/2, \tag{3.16}$$

where $\mathbf{W}_j$ is $n \times n$ diagonal matrix with diagonal elements $(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{nj}(\boldsymbol{\Psi}^{(r)}))$ obtained from (3.4). Applying (3.16), $\widehat{\boldsymbol{\beta}}_j^{(r+1)}; j = 1, \ldots, M$ is updated by

$$\widehat{\boldsymbol{\beta}}_{R,j}^{(r+1)} = \left(\mathbf{X}^\top\mathbf{W}_j\mathbf{X} + k_j\mathbb{I}\right)^{-1}\mathbf{X}^\top\mathbf{W}_j\mathbf{y}. \tag{3.17}$$

Lemma 3.1. *Under the assumptions of mixture of regression models (3.1), suppose* $\lambda_{1j}, \ldots, \lambda_{pj}$ *and* $u_{1j}, \ldots, u_{pj}$ *be eigenvalues and orthonormal eigenvectors of* $\mathbf{X}^\top \mathbf{W}_j \mathbf{X}$ *where* $\mathbf{W}_j$ *is* $n \times n$ *diagonal matrix with entries* $(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{nj}(\boldsymbol{\Psi}^{(r)}))$ *under Ridge EM algorithm. Let* $\boldsymbol{\Lambda}_j = diag(\lambda_{1j}, \ldots, \lambda_{pj})$ *and* $\mathbf{U}_j = [u_{1j}, \ldots, u_{pj}]$. *Then the canonical weighted Ridge estimator in each component regression is given by*

$$\widehat{\boldsymbol{\alpha}}_{R,j} = (\boldsymbol{\Lambda}_j + k_j \mathbb{I})^{-1} \boldsymbol{\Lambda}_j^{1/2} \mathbf{V}_1^\top \mathbf{W}_j^{1/2} \mathbf{y}.$$

and

$$\widehat{\boldsymbol{\beta}}_{R,j} = \mathbf{U}_j \widehat{\boldsymbol{\alpha}}_{R,j}$$

with $\mathbf{V}_1 = [v_{1j}, \ldots, v_{pj}]$ *where* $v_{1j}, \ldots, v_{pj}$ *are the orthonormal eigenvectors of* $\mathbf{W}_j^{1/2} \mathbf{X}\mathbf{X}^\top \mathbf{W}_j^{1/2}$.

Proof: The positive eigenvalues of $\mathbf{X}^\top \mathbf{W}_j \mathbf{X}$ and $\mathbf{W}_j^{1/2} \mathbf{X}\mathbf{X}^\top \mathbf{W}_j^{1/2}$ must be the same. Hence, the eigenvalues of $\mathbf{W}_j^{1/2} \mathbf{X}\mathbf{X}^\top \mathbf{W}_j^{1/2}$ are given by $\lambda_{1j}, \ldots, \lambda_{pj}$ and the other $(n - p)$ values must be zero. From singular value decomposition, it is easy to see $\mathbf{V}_1 = \mathbf{W}_j^{1/2} \mathbf{X} \mathbf{U}_j \boldsymbol{\Lambda}_j^{-1/2}$ and $\boldsymbol{\Lambda}_j^{1/2} = \mathbf{V}_1^\top \mathbf{W}_j^{1/2} \mathbf{X} \mathbf{U}_j$. From the definition of $\mathbf{V}_1$ and $\boldsymbol{\Lambda}_j^{1/2}$, we can show

$$\mathbf{W}_j^{1/2} \mathbf{X} = \mathbf{V}_1 \mathbf{V}_1^\top \mathbf{W}_j^{1/2} \mathbf{X} \mathbf{U}_j \mathbf{U}_j^\top = \mathbf{V}_1 \boldsymbol{\Lambda}_j^{1/2} \mathbf{U}_j^\top. \tag{3.18}$$

From (3.18), we can write the canonical form of the regression by

$$\mathbf{W}_j^{1/2} \mathbf{y} = \mathbf{W}_j^{1/2} \mathbf{X} \boldsymbol{\beta}_j + \mathbf{W}_j^{1/2} \epsilon = \mathbf{V}_1 \boldsymbol{\Lambda}_j^{1/2} \mathbf{U}_j^\top \boldsymbol{\beta}_j + \mathbf{W}_j^{1/2} \epsilon = \mathbf{V}_1 \boldsymbol{\Lambda}_j^{1/2} \boldsymbol{\alpha}_j + \mathbf{W}_j^{1/2} \epsilon. \tag{3.19}$$

From (3.19), we can derive the canonical form of the weighted Ridge estimator in

each component by

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{R,j} &= \left((\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2})^\top(\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}) + k_j\mathbb{I}\right)^{-1}(\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2})^\top\mathbf{W}_j^{1/2}\mathbf{y} \\
&= \left(\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2} + k_j\mathbb{I}\right)^{-1}\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y} \\
&= (\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y}.
\end{aligned}$$

$$\begin{aligned}
\mathbf{U}_j\widehat{\boldsymbol{\alpha}}_{R,j} =& \mathbf{U}_j(\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}\mathbf{U}_j^\top\mathbf{U}_j\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y} \\
=& \left(\mathbf{U}_j\boldsymbol{\Lambda}_j^{1/2}\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top + k_j\mathbb{I}\right)^{-1}(\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top)^\top\mathbf{W}_j^{1/2}\mathbf{y} \\
=& \left((\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top)^\top(\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top) + k_j\mathbb{I}\right)^{-1}(\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top)^\top\mathbf{W}_j^{1/2}\mathbf{y} \\
=& \left((\mathbf{W}_j^{1/2}\mathbf{X})^\top(\mathbf{W}_j^{1/2}\mathbf{X}) + k_j\mathbb{I}\right)^{-1}(\mathbf{W}_j^{1/2}\mathbf{X})^\top\mathbf{W}_j^{1/2}\mathbf{y} \\
=& \left(\mathbf{X}^\top\mathbf{W}_j\mathbf{X} + k_j\mathbb{I}\right)^{-1}\mathbf{X}^\top\mathbf{W}_j\mathbf{y}
\end{aligned}$$

□

From Ridge weighted least square (3.16), the variance term can be updated by

$$\widehat{\sigma}_{R,j}^{2(r+1)} = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R^{(r+1)})^\top\mathbf{W}_j^{(r)}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\boldsymbol{\Psi}^{(r)})}, \tag{3.20}$$

where $\widehat{\boldsymbol{\beta}}_R^{(r+1)} = (\widehat{\boldsymbol{\beta}}_{R,1}^{(r+1)}, \ldots, \widehat{\boldsymbol{\beta}}_{R,M}^{(r+1)})$. There are various methods available in the literature for estimation of k_j. Following Hoerl and Kennard (1970) and Liu (2003), we estimate the parameter by $\widehat{k_j} = p\widehat{\sigma}_{ML,j}^2/\widehat{\boldsymbol{\beta}}_{ML,j}^\top\widehat{\boldsymbol{\beta}}_{ML,j}$ where $\widehat{\sigma}_{ML,j}^2$ and $\widehat{\boldsymbol{\beta}}_{ML,j}$ are calculated from (3.11) and (3.10), respectively. The E- and M-steps are repeatedly computed until $|\ell^R(\boldsymbol{\Psi}^{(r+1)}) - \ell^R(\boldsymbol{\Psi}^{(r)})| < \epsilon$, where ϵ is a user defined tolerance.

3.4.1 Ridge CEM Algorithm

In a mixture of regressions, the CEM technique can also be used to obtain the Ridge estimates of the parameters. We must accommodate a C-step between E- and M-steps in the Ridge EM algorithm, similar to the CEM algorithm given in subsection 3.3. Here the E-step remains the same as before. Like the C-step of ML method, we classify the observations to partitions $\mathbf{P}^{(r+1)} = (P_1^{(r+1)}, \ldots, P_M^{(r+1)})$ based on the maximum probability of memberships; that is

$$P_j^{(r+1)} = \{(\mathbf{x}_i, y_i); \tau_{ij}(\boldsymbol{\Psi}^{(r)}) = \arg\max_h \tau_{ih}(\boldsymbol{\Psi}^{(r)})\}, \quad \forall j = 1, \ldots, M.$$

Based on $\mathbf{P}^{(r+1)}$, we use (3.12) to update the mixing proportions of the mixture. The Ridge parameters are calculated using a method similar to the Ridge EM technique. We apply the Ridge weighted least square (3.16) to each partition $P_j^{(r+1)}$ and update the coefficients and variance term of each component regression by

$$\widehat{\boldsymbol{\beta}}_{R,j}^{(r+1)} = \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j + k_j \mathbb{I}\right)^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{y}_j, \tag{3.21}$$

$$\widehat{\sigma}_{R,j}^{2(r+1)} = \frac{(\mathbf{y}_j - \mathbf{X}_j \widehat{\boldsymbol{\beta}}_R^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y}_j - \mathbf{X}_j \widehat{\boldsymbol{\beta}}_R^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\boldsymbol{\Psi}^{(r)})}, \tag{3.22}$$

where $\mathbf{X}_j$ is $(n_j \times p)$ design matrix and $\mathbf{y}_j$ is vector of responses from observations classified to $P_j^{(r+1)}$. $\mathbf{W}_j^{(r)}$ is the diagonal weight matrix with entries $\left(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{n_j,j}(\boldsymbol{\Psi}^{(r)})\right)$ from (3.4). Finally, the E-, C- and M- steps under Ridge estimation procedure are alternated until convergence criterion is statified.

3.4.2 Ridge SEM Algorithm

The stochastic EM algorithm can be used to implement the Ridge estimation method. The S-step, like the SEM of the ML technique, determines the component membership of observations under the Ridge approach stochastically by $\mathbf{Z}_i^* = (Z_{i1}^*, \ldots, Z_M^*) \overset{iid}{\sim} \text{Multi}(1, \tau_{i1}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{iM}(\boldsymbol{\Psi}^{(r)})); i = 1, \ldots, n$ and updates $\mathbf{P}^{(r+1)} = (P_1^{(r+1)}, \ldots, P_M^{(r+1)})$ such that $P_j^{(r+1)} = \{(\mathbf{x}_i, y_i); Z_{ij}^* = 1\}; \forall j = 1, \ldots, M$. Based on this stochastic partition of S-step, we update the mixture parameters by (3.12), (3.21) and (3.22).

Lemma 3.2. *Under the assumptions of mixture of regression models (3.1), with component regression models* $\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \epsilon$ *based on* n_j *observations with rank*$(\mathbf{X}_j) = p$. *Suppose* $\lambda_{1j}, \ldots, \lambda_{pj}$ *and* $u_{1j}, \ldots, u_{pj}$ *be eigenvalues and orthonormal eigenvectors of* $\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j$ *where* $\mathbf{W}_j$ *is* $n_j \times n_j$ *diagonal matrix with entries* $(\tau_{ij}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{nj}(\boldsymbol{\Psi}^{(r)}))$ *under Ridge CEM or Ridge SEM algorithm. Let* $\boldsymbol{\Lambda}_j = diag(\lambda_{1j}, \ldots, \lambda_{pj})$ *and* $\mathbf{U}_j = [u_{1j}, \ldots, u_{pj}]$. *Then The canonical weighted Ridge estimator in each component regression is given by*

$$\widehat{\boldsymbol{\alpha}}_{R,j} = (\boldsymbol{\Lambda}_j + k_j \mathbb{I})^{-1} \boldsymbol{\Lambda}_j^{1/2} \mathbf{V}_1^\top \mathbf{W}_j^{1/2} \mathbf{y}_j.$$

and

$$\widehat{\boldsymbol{\beta}}_{R,j} = \mathbf{U}_j \widehat{\boldsymbol{\alpha}}_{R,j},$$

with $\mathbf{V}_1 = [v_{1j}, \ldots, v_{pj}]$ *where* $v_{1j}, \ldots, v_{pj}$ *are the orthonormal eigenvectors of* $\mathbf{W}_j^{1/2} \mathbf{X}_j \mathbf{X}_j^\top \mathbf{W}_j^{1/2}$.

Proof: The lemma can be proved in a similar vein to Lemma 3.1. □

Finally, the E-, S- and M-steps are iterated until either stopping rule is satisfied or algorithm reaches a pre-specified number iterations.

3.5 Liu-type Estimation Method

When the design matrix is severely ill-conditioned, using the Ridge estimator to add small values to the diagonal members may not be enough to solve the multicollinearity problem. On the other hand, increasing the Ridge parameter may result in a more extensive bias in the Ridge estimation approach. Liu (2003) proposed the Liu-type (LT) shrinkage approach for estimating regression parameters when there is severe multicollinearity. Like the Ridge method, the LT approach optimizes the estimating equation while applying the LT penalty to control the multicollinearity problem. The LT penalty is given by

$$(-\frac{d}{k^{1/2}})\widehat{\boldsymbol{\beta}} = k^{1/2}\boldsymbol{\beta} + \epsilon', \tag{3.23}$$

where $\widehat{\boldsymbol{\beta}}$ can be any estimator of coefficients and $d \in \mathbb{R}$ and $k > 0$ are two tuning parameters of the LT estimation method. In this section, we create the LT shrinkage approach for estimating the unknown parameters of the mixture of regression models (3.1). By maximizing the log-likelihood function (3.2) subject to the LT penalty, we find the LT estimate of $\boldsymbol{\Psi}$.

The penalized log-likelihood function based on the observed data is not tractable concerning the component parameters. We apply an unsupervised technique to design the LT estimation procedure and use the EM algorithm to estimate the unknown parameters of the mixture model iteratively. First we have to introduce latent vectors $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ in order to represent the component membership

of the i-th observation $(\mathbf{x}_i, y_i); i = 1, \ldots, n$. Let $(\mathbf{X}, \mathbf{y}, \mathbf{Z})$ represent the complete data. Then EM algorithm under the LT method proceeds as follows.

On $(r+1)$-th iteration, the E-step stays similar to the E-step under ML approach. The mixing proportion of the model under LT estimation is updated by (3.8). From (3.7) we have to maximize $\mathbf{Q}_2(\boldsymbol{\theta}, \boldsymbol{\Psi}^{(r)})$ under the LT penalty within each component to estimate the regression parameters. The LT penalized log-likelihood function can be written as a weighted least square constrained on LT penalty as

$$\widehat{\boldsymbol{\beta}}_{LT,j}^{(r+1)} = \arg\min_{\boldsymbol{\beta}_j} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W}_j (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \left[(-\frac{d_j}{k_j^{1/2}}) \widehat{\boldsymbol{\beta}}_j - k_j^{1/2} \boldsymbol{\beta}_j \right]^\top \left[(-\frac{d_j}{k_j^{1/2}}) \widehat{\boldsymbol{\beta}}_j - k_j^{1/2} \boldsymbol{\beta}_j \right], \tag{3.24}$$

where $\widehat{\boldsymbol{\beta}}_j$ can be any coefficient estimate and $\mathbf{W}_j$ is a weight diagonal matrix with diagonal elements $(\tau_{1j}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{nj}(\boldsymbol{\Psi}^{(r)}); j = 1, \ldots, M$. From (3.24), the coefficients and variance term in each component regression are updated by

$$\widehat{\boldsymbol{\beta}}_{LT,j}^{(r+1)} = \left(\mathbf{X}^\top \mathbf{W}_j \mathbf{X} + k_j \mathbb{I}\right)^{-1} (\mathbf{X}^\top \mathbf{W}_j \mathbf{y} - d_j \widehat{\boldsymbol{\beta}}_j), \tag{3.25}$$

$$\widehat{\sigma}_{LT,j}^{2(r+1)} = \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{LT}^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{LT}^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\boldsymbol{\Psi}^{(r)})}, \tag{3.26}$$

where $\widehat{\boldsymbol{\beta}}_{LT}^{(r+1)} = (\widehat{\boldsymbol{\beta}}_{LT,1}^{(r+1)}, \ldots, \widehat{\boldsymbol{\beta}}_{LT,J}^{(r+1)})$. In order to estimates (3.25) and (3.26), first we have to estimate the LT parameters (k_j, d_j) for each component regression. From Liu (2003), we can estimate k_j in the j-th component by $\widehat{k}_{LT,j} = (\lambda_{1,j} - 100\lambda_{p,j})/99$ where $\lambda_{1,j}$ and $\lambda_{p,j}$ are maximum and minimum eigenvalues of $\mathbf{X}^\top \mathbf{W}_j \mathbf{X}$ on the $(r+1)$-the iteration of the EM algorithm.

Lemma 3.3. *Under the assumptions of Lemma (3.1), the canonical LT estimate in the j-th component regression $j = 1, \ldots, M$ under EM algorithm is given by*

$$\widehat{\boldsymbol{\alpha}}_{LT,j} = (\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}\,(\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\widehat{\boldsymbol{\beta}}_j),$$

and

$$\widehat{\boldsymbol{\beta}}_{LT,j} = \mathbf{U}_j\widehat{\boldsymbol{\alpha}}_{LT,j},$$

where $\widehat{\boldsymbol{\alpha}}_j$ is the canonical estimate of $\boldsymbol{\beta}_j$ and $\mathbf{V}_1 = [v_{1j}, \ldots, v_{pj}]$ with $v_{1j}, \ldots, v_{pj}$ are the orthonormal eigenvectors of $\mathbf{W}_j^{1/2}\mathbf{X}\mathbf{X}^\top\mathbf{W}_j^{1/2}$.

Proof: From Lemma 3.1, (3.18) and (3.19), we can write the canonical form of the weighted LT estimator in each component regression as

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{LT,j} &= \left((\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2})^\top(\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}) + k_j\mathbb{I}\right)^{-1}\left((\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2})^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\widehat{\boldsymbol{\alpha}}_j\right)\\
&= \left(\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2} + k_j\mathbb{I}\right)^{-1}\left(\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\widehat{\boldsymbol{\alpha}}_j\right)\\
&= (\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}\left(\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\widehat{\boldsymbol{\alpha}}_j\right).
\end{aligned}$$

$$\begin{aligned}
\mathbf{U}_j\widehat{\boldsymbol{\alpha}}_{R,j} =& \mathbf{U}_j\,(\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}\left(\boldsymbol{\Lambda}_j^{1/2}\mathbf{V}_1^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\widehat{\boldsymbol{\alpha}}_j\right)\\
&= \left(\mathbf{U}_j\boldsymbol{\Lambda}_j^{1/2}\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top + k_j\mathbb{I}\right)^{-1}\left((\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top)^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\mathbf{U}_j\widehat{\boldsymbol{\alpha}}_j\right)\\
&= \left((\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top)^\top(\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top) + k_j\mathbb{I}\right)^{-1}\left((\mathbf{V}_1\boldsymbol{\Lambda}_j^{1/2}\mathbf{U}_j^\top)^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\widehat{\boldsymbol{\beta}}_j\right)\\
&= \left((\mathbf{W}_j^{1/2}\mathbf{X})^\top(\mathbf{W}_j^{1/2}\mathbf{X}) + k_j\mathbb{I}\right)^{-1}\left((\mathbf{W}_j^{1/2}\mathbf{X})^\top\mathbf{W}_j^{1/2}\mathbf{y} - d_j\widehat{\boldsymbol{\beta}}_j\right)\\
&= \left(\mathbf{X}^\top\mathbf{W}_j\mathbf{X} + k_j\mathbb{I}\right)^{-1}\left(\mathbf{X}^\top\mathbf{W}_j\mathbf{y} - d_j\widehat{\boldsymbol{\beta}}_j\right).
\end{aligned}$$

□

Following Liu (2003) and Lemma 3.3, the optimal d_j can be obtained by the next lemma within each component of the mixture of regression models.

Lemma 3.4. *Under the assumptions of Lemma 3.1 for $k_j > 0$ and $\widehat{\boldsymbol{\alpha}}_j = \widehat{\boldsymbol{\alpha}}_{ML,j}$,*

$$d_j = \sum_{m=1}^{p} \left(\lambda_{mj}(\sigma_j^2 - k_j\alpha_{mj}^2)/(\lambda_{mj}+k_j)^3\right) / \sum_{m=1}^{p} \left(\left(\lambda_{mj}(\lambda_{mj}\alpha_{mj}^2 + \sigma_j^2)\right)/(\lambda_{mj}+k_j)^4\right)$$

minimizes the MSE($\widehat{\boldsymbol{\alpha}}_{LT,j}$) within each component of the mixture (3.1) in the EM algorithm under LT method.

Proof: Since $\widehat{\boldsymbol{\alpha}}_{ML,j} = \boldsymbol{\Lambda}_j^{-1/2}\mathbf{V}_1^\top \mathbf{W}_j^{1/2}\mathbf{y}$, it is easy to show that

$$\widehat{\boldsymbol{\alpha}}_{LT,j} = (\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}(\boldsymbol{\Lambda}_j - d_j\mathbb{I})\widehat{\boldsymbol{\alpha}}_{ML,j}. \tag{3.27}$$

From (3.27), the bias and covariance of $\widehat{\boldsymbol{\alpha}}_{LT,j}$ are computed by

$$\begin{aligned}
\text{Bias}(\widehat{\boldsymbol{\alpha}}_{LT,j}) &= \mathbb{E}(\widehat{\boldsymbol{\alpha}}_{LT,j}) - \boldsymbol{\alpha}_j \\
&= (\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}(\boldsymbol{\Lambda}_j - d_j\mathbb{I})\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_j \\
&= -(\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}(k_j + d_j)\boldsymbol{\alpha}_j.
\end{aligned}$$

$$\text{cov}(\widehat{\boldsymbol{\alpha}}_{LT,j}) = \sigma_j^2(\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}(\boldsymbol{\Lambda}_j - d_j\mathbb{I})\boldsymbol{\Lambda}_j^{-1}(\boldsymbol{\Lambda}_j - d_j\mathbb{I})(\boldsymbol{\Lambda}_j + k_j\mathbb{I})^{-1}.$$

Following Liu (2003), we can find the MSE($\widehat{\boldsymbol{\alpha}}_{LT,j}$) using the bias and covariance as follows

$$\begin{aligned}
\text{MSE}(\widehat{\boldsymbol{\alpha}}_{LT,j}) &= ||\text{Bias}(\widehat{\boldsymbol{\alpha}}_{LT,j})||^2 + \text{tr}(\text{cov}(\widehat{\boldsymbol{\alpha}}_{LT,j})) \\
&= \sum_{m=1}^{p}(d_j + k_j)^2\alpha_m^2/(\lambda_m + k_j)^2 + \sigma_j^2\sum_{m=1}^{p}(d_j - \lambda_m)^2/\lambda_m(\lambda_m + k_j)^2.
\end{aligned}$$

Differentiating MSE($\widehat{\boldsymbol{\alpha}}_{LT,j}$) with resect to d_j, it is easy to obtain

$$d_{opt,j}\sum_{m=1}^{p}\left((\lambda_m\alpha_m^2+\sigma_j^2)/\lambda_m(\lambda_m+k_j)^2\right)=\sum_{m=1}^{p}\left((\sigma_j^2-k_j\alpha_m^2)/(\lambda_m+k_j)^2\right).$$

□

Despite the fact that Lemma 3.4 leads the way to estimating the best LT parameter d_j within each component regression of the EM method, the optimal value is still dependent on unknown quantities such as σ_j, k_j, $\boldsymbol{\alpha}_j$, and $\lambda_{m,j}$ for $m=1,\ldots,p$ and $j=1,\ldots,M$. From Lemma 3.4, we propose a practical approach where $d_j, j=1,\ldots,M$ can be updated in the $(r+1)$-the iteration of the EM algorithm by

$$\widehat{d}_j=\sum_{m=1}^{p}\left(\lambda_{mj}(\widehat{\sigma}_{R,j}^2-\widehat{k}_j\widehat{\alpha}_{R,mj}^2)/(\lambda_{mj}+\widehat{k}_j)^3\right)/\sum_{m=1}^{p}\left(\left(\lambda_{mj}(\lambda_{mj}\widehat{\alpha}_{R,mj}^2+\widehat{\sigma}_j^2)\right)/(\lambda_{mj}+\widehat{k}_j)^4\right), \tag{3.28}$$

where $\widehat{k}_j=\widehat{k}_{LT,j}$, $\widehat{\alpha}_{R,j}=(\widehat{\alpha}_{R,1j},\ldots,\widehat{\alpha}_{R,pj})$ is given by Lemma 3.1 and $(\lambda_{1j},\ldots,\lambda_{pj})$ are eigenvalues of $\mathbf{X}_j^\top\mathbf{W}_j\mathbf{X}_j$ with $\widehat{\sigma}_{R,j}^2$ from (3.20). Until the stopping requirement is met, the E- and M-steps are alternated. The proposed LT estimation method is now known as iterative LT because the parameters k_j and d_j are changed in each iteration of the EM algorithm.

Unlike iterative LT method, in order to estimate the LT parameters based on Ridge estimates $\widehat{\sigma}_{R,j}$ and $\widehat{\boldsymbol{\beta}}_{R,j}$, we can follow Hoerl et al. (1975). In other words, the EM algorithm iteratively updates the mixture parameter $\boldsymbol{\Psi}$, while the k_j parameter is only estimated once during the EM algorithm, using the final Ridge estimates from Subsection 3.4. Here, we estimate the parameters by $\widehat{k}_{LT,j}=p\widehat{\sigma}_{R,j}/\widehat{\boldsymbol{\beta}}_{R,j}^\top\widehat{\boldsymbol{\beta}}_{R,j}$ and $\widehat{d}_j$

from (3.28) where $\widehat{\boldsymbol{\beta}}_{R,j}$ and $\widehat{\sigma}_{R,j}$ are obtained from (3.17) and (3.20), respectively. This LT estimation method is henceforth is called HKP Liu-type.

3.5.1 Liu-type CEM Algorithm

The CEM algorithm partitions the observations in C-step, as in previous subsections, and then updates the unknown parameters within each partition. The E-step stays the same on the $(r+1)$-th iteration of the CEM algorithm. The C-step classifies the observations into partition $\mathbf{P}^{(r+1)} = (P_1^{(r+1)}, \ldots, P_M^{(r+1)})$ where $P_j^{(r+1)} = \{(\mathbf{x}_i, y_i); \tau_{ij}(\boldsymbol{\Psi}^{(r)}) = \arg\max_h \tau_{ih}(\boldsymbol{\Psi}^{(r)})\}$ with $(\tau_{i1}(\boldsymbol{\Psi}^{(r)}, \ldots, \tau_{iM}(\boldsymbol{\Psi}^{(r)})$ are obtained from (3.4). Using $\mathbf{P}^{(r+1)}$, we update the mixing proportions from (3.12). The LT tuning parameters (k_j, d_j) must then be estimated in each iteration of the CEM algorithm, similar to the Liu-type EM algorithm. we propose $\widehat{k}_{LT,j} = (\lambda_{1,j} - 100\lambda_{p,j})/99$ where $\lambda_{1,j}$ and $\lambda_{p,j}$ are maximum and minimum eigenvalues of $\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j$.

Lemma 3.5. *Under the assumptions of Lemma (3.1), the canonical LT estimate in the j-th component regression $j = 1, \ldots, M$ under CEM algorithm is given by*

$$\widehat{\boldsymbol{\alpha}}_{LT,j} = (\boldsymbol{\Lambda}_j + k_j)^{-1} (\boldsymbol{\Lambda}_j^{1/2} \mathbf{V}_1^\top \mathbf{W}_j^{1/2} \mathbf{y} - d_j \widehat{\boldsymbol{\alpha}}_j),$$

and

$$\widehat{\boldsymbol{\beta}}_{LT,j} = \mathbf{U}_j \widehat{\boldsymbol{\alpha}}_{LT,j},$$

where $\widehat{\boldsymbol{\alpha}}_j$ is the canonical estimate of $\boldsymbol{\beta}_j$ and $\mathbf{V}_1 = [v_{1j}, \ldots, v_{pj}]$ with $v_{1j}, \ldots, v_{pj}$ are the orthonormal eigenvectors of $\mathbf{W}_j^{1/2} \mathbf{X}_j \mathbf{X}_j^\top \mathbf{W}_j^{1/2}$.

Proof: The lemma can be proved in a similar vein to Lemma 3.3. □

From Lemma 3.5 and Lemma 3.4, one can estimate parameter d_j based on partition $P_j^{(r+1)}$ from (3.28) where $(\lambda_{1j}, \ldots, \lambda_{pj})$ are eigenvalues of $\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j$ and $\widehat{\sigma}_{R,j}^2$ from (3.22). To estimate the regression parameters, we implement a weighted least square based on LT penalty as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{LT,j}^{(r+1)} &= \arg\min_{\boldsymbol{\beta}_j} (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j)^\top \mathbf{W}_j (\mathbf{y}_j - \mathbf{X}_j\boldsymbol{\beta}_j) \\ &+ \left[(-\frac{d_j}{k_{LT,j}^{1/2}})\widehat{\boldsymbol{\beta}}_j - k_{LT,j}^{1/2}\boldsymbol{\beta}_j\right]^\top \left[(-\frac{d_j}{k_{LT,j}^{1/2}})\widehat{\boldsymbol{\beta}}_j - k_{LT,j}^{1/2}\boldsymbol{\beta}_j\right],\end{aligned} \tag{3.29}$$

where $\mathbf{y}_j$ and $\mathbf{X}_j$ are response vector and design matrix under $P_j^{(r+1)}$ and $\widehat{\boldsymbol{\beta}}_j$ can be any estimate for $\boldsymbol{\beta}_j$. Also, $\mathbf{W}_j$ is a weight diagonal matrix with diagonal elements $(\tau_{1j}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{n_j,j}(\boldsymbol{\Psi}^{(r)}); j = 1, \ldots, M$. One can easily find the solution to (3.24) and update the regression parameters by

$$\widehat{\boldsymbol{\beta}}_{LT,j}^{(r+1)} = \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j + k_{LT,j}\mathbb{I}\right)^{-1} (\mathbf{X}_j^\top \mathbf{W}_j \mathbf{y}_j - d_j \widehat{\boldsymbol{\beta}}_j), \tag{3.30}$$

$$\widehat{\sigma}_{LT,j}^{2(r+1)} = \frac{(\mathbf{y}_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}}_{LT}^{(r+1)})^\top \mathbf{W}_j^{(r)} (\mathbf{y}_j - \mathbf{X}_j\widehat{\boldsymbol{\beta}}_{LT}^{(r+1)})}{\sum_{i=1}^n \tau_{ij}(\boldsymbol{\Psi}^{(r)})}, \tag{3.31}$$

with $\widehat{\boldsymbol{\beta}}_{LT}^{(r+1)} = (\widehat{\boldsymbol{\beta}}_{LT,1}^{(r+1)}, \ldots, \widehat{\boldsymbol{\beta}}_{LT,J}^{(r+1)})$. The E-, C- and M-steps are repeatedly computed until the convergence criterion is satisfied.

Unlike the iterative Liu-type CEM algorithm, the HKP Liu-type CEM method can estimate the parameters of the mixture model since the LT tuning parameter

$k_j, j = 1, \ldots, M$ is only updated once during the process. From Hoerl et al. (1975) and Liu (2003) , we propose to estimate $\widehat{k}_{LT,j} = p\widehat{\sigma}_{R,j}/\widehat{\boldsymbol{\beta}}_{R,j}^{\top}\widehat{\boldsymbol{\beta}}_{R,j}$ and $\widehat{d}_j$ from (3.28) where $\widehat{\boldsymbol{\beta}}_{R,j}$ and $\widehat{\sigma}_{R,j}$ come from (3.21) and (3.22), respectively.

3.5.2 Liu-type SEM Algorithm

Similar to the SEM algorithms that described earlier, the S-step parition the observations stochastically using Multi$(1, \tau_{i1}(\boldsymbol{\Psi}^{(r)}), \ldots, \tau_{iM}(\boldsymbol{\Psi}^{(r)}))$ for $i = 1, \ldots, n$. Once the partition is established, the rest of the Liu-type SEM estimation method is computed in a similar vein to the Liu-type CEM algorithm. The LT tuning parameters (k_j, d_j) must then be estimated in each iteration of the SEM algorithm, similar to the Liu-type EM and CEM algorithms.

Chapter 4

Shrinkage Estimators for Mixture of Logistic Regressions

This chapter focuses on the finite mixture of logistic regression models. This model is merely a generalization of logistic regression when the observed data come from various M components. Multicollinearity significantly impacts the maximum likelihood (ML) estimates of a mixture of logistic regressions, just as it does on logistic regression. We developed the Liu-type (LT) shrinkage estimator for the mixture of logistic regression models (Ghanem et al., 2022a).

This chapter is organized as follows. Section 4.2.1 describes the ML estimation method. Sections 4.3 and 4.4 describe the Ridge and LT methods in estimating the parameters of the mixture of logistic regression models.

4.1 Introduction

One of the most vital families of the mixture model is the mixture of logistic regression models. A finite mixture of the logistic regression model was applied to analyze

the heterogeneity within the merging population. This model can automatically show important hidden information regarding the population's characteristics. The EM algorithm and Newton-Raphson algorithm were used to estimate the parameters. Dempster et al. (1977) used the expectation-maximization (EM) approach to get the ML estimates of FMMs and a mixture of logistic regression models. Murray (1999a,b) and Wang and Puterman (1998) applied the ML approach to estimate the parameters of a finite mixture of logistic regression models. Mixture models have many applications in the core of statistical sciences, such as data classification and modeling from various sampling structures, stratified sampling (Wedel et al., 1998) and ranked set sampling (Hatefi et al., 2015, 2020).

4.2 Statistical Methods

The mixture of logistic regression is a generalization of logistic regression, when the observed data come from different components. Let M represents the number of the components of the mixture of logistic regression models. While the number of components M is assumed to be known, the problem is addressed as an unsupervised learning approach when the component membership of observations is unknown and should be estimated. From Subsection 2.10, the log-likelihood of the mixture of logistic regression models follows

$$\ell(\mathbf{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{M} \pi_j [p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{y_i} [1 - p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{(1-y_i)} \right\}, \tag{4.1}$$

where

$$p_j(\mathbf{x}_j; \boldsymbol{\beta}_j) = g^{-1}(\mathbf{x}_i; \boldsymbol{\beta}_j), \tag{4.2}$$

and $\pi = (\pi_1, \ldots, \pi_M)$ represent the vector of the mixing proportions with $\pi_j > 0$ and $\sum_{j=1}^{M} \pi_j = 1$. Also, we use $\boldsymbol{\Psi} = (\pi, \boldsymbol{\beta})$ with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$ to represent the vector of all unknown parameters of the mixture of logistic regression models.

4.2.1 ML Estimation Method

In estimating the parameters of the mixture model, there is no closed form for the maximizer of the log-likelihood function (4.1). As a result, we consider $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ to be incomplete data and propose an expectation-maximization (EM) approach to derive an ML estimate of $\boldsymbol{\Psi}$. Suppose $\{(\mathbf{x}_i, y_i, \mathbf{Z}_i), i = 1, \ldots, n\}$ denote the complete data where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM})$ is the latent variable repressing the component membership of the i-th subject with

$$Z_{ij} = \begin{cases} 1 & \text{if the } i\text{-th subject comes from the } j\text{-th component,} \\ 0 & o.w., \end{cases}$$

Given that $\mathbf{Z}_i \overset{iid}{\sim} \text{Multi}(1, \pi_1, \ldots, \pi_M)$, the joint distribution of $(y_i, \mathbf{Z}_i)$ can be written as

$$f(y_i, \mathbf{z}_i) = \prod_{j=1}^{M} \left\{ \pi_j [p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{y_i} [1 - p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{(1-y_i)} \right\}^{z_{ij}}. \tag{4.3}$$

From above, it is easy to show $\mathbf{Z}_i | y_i \overset{iid}{\sim} \text{Multi}(1, \tau_{i1}(\boldsymbol{\Psi}), \ldots, \tau_{iM}(\boldsymbol{\Psi}))$ where

$$\tau_{ij}(\boldsymbol{\Psi}) = \frac{\pi_j [p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{y_i} [1 - p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{(1-y_i)}}{\sum_{j=1}^{M} \pi_j [p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{y_i} [1 - p_j(\mathbf{x}_i; \boldsymbol{\beta}_j)]^{(1-y_i)}}. \tag{4.4}$$

By using the latent variables $\mathbf{Z}_i$, the complete log-likelihood function of $\boldsymbol{\Psi}$ is

given by

$$\ell_c(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\log(\pi_j) + \sum_{i=1}^{n}\sum_{j=1}^{M} z_{ij}\log\left\{[p_j(\mathbf{x}_i;\boldsymbol{\beta}_j)]^{y_i}[1-p_j(\mathbf{x}_i;\boldsymbol{\beta}_j)]^{(1-y_i)}\right\}. \tag{4.5}$$

The EM algorithm breaks down the estimating process into two iterative steps: expectation (E-step) and maximization (M-step). Unlike Chapter 3 in the analysis of a mixture of linear regressions, we only propose stochastic EM (SEM) of (Celeux, 1985) algorithm to estimate the mixing proportions and coefficients of the mixture of logistic regression in Chapter 4. The SEM algorithm is a redesigned version of the EM method that includes a stochastic classification step (S-step) between the E- and M-steps.

Like an iterative approach, the SEM algorithm requires initial values to begin the estimating process. Let $\boldsymbol{\Psi}^{(0)} = (\pi^{(0)}, \boldsymbol{\beta}^{(0)})$ denote the starting points of algorithm. We discuss how the E, S, and M steps are implemented in the $(l+1)$-th iteration, where $\boldsymbol{\Psi}(l)$ represents the update from the l-th iteration, to better comprehend the SEM algorithm.

E-Step: First the conditional expectational of latent variables must be computed given incomplete data. Hence,

$$\mathbb{E}_{\boldsymbol{\Psi}^{(l)}}(Z_{ij}|y_i) = \tau_{ij}(\boldsymbol{\Psi})|_{\boldsymbol{\Psi}=\boldsymbol{\Psi}^{(l)}} = \tau_{ij}(\boldsymbol{\Psi}^{(l)}),$$

where $\tau_{ij}(\boldsymbol{\Psi}^{(l)})$ is calculated from (4.4). The conditional expectation of the log-likelihood function (4.5) can be re-written by

$$\mathbf{Q}(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(l)}) = \mathbb{E}_{\boldsymbol{\Psi}^{(l)}}(\ell_c(\boldsymbol{\beta})|\mathbf{y}, \boldsymbol{\Psi}^{(l)}) = \mathbf{Q}_1(\pi, \boldsymbol{\Psi}^{(l)}) + \mathbf{Q}_2(\boldsymbol{\beta}, \boldsymbol{\Psi}^{(l)}),$$

where

$$\mathbf{Q}_1(\pi, \mathbf{\Psi}^{(l)}) = \sum_{i=1}^{n}\sum_{j=1}^{M} \tau_{ij}(\mathbf{\Psi}^{(l)}) \log(\pi_j), \tag{4.6}$$

and

$$\mathbf{Q}_2(\xi, \mathbf{\Psi}^{(l)}) = \sum_{i=1}^{n}\sum_{j=1}^{M} \tau_{ij}(\mathbf{\Psi}^{(l)}) \log\left\{[p_j(\mathbf{x}_i;\boldsymbol{\beta}_j)]^{y_i}[1-p_j(\mathbf{x}_i;\boldsymbol{\beta}_j)]^{(1-y_i)}\right\}. \tag{4.7}$$

S-Step: We partition the subjects into $\mathbf{P}^{(l+1)} = (P_1^{(l+1)}, \ldots, P_M^{(l+1)})$ based on a stochastic assignment $(Z_{i1}^*, \ldots, Z_{iM}^*)$, given their posterior probability memberships $(\tau_{i1}(\mathbf{\Psi}^{(l)}), \ldots, \tau_{iM}(\mathbf{\Psi}^{(l)}))$. In other words, we generate $\mathbf{Z}_i^* \overset{iid}{\sim} \text{Multi}(1, \tau_{i1}(\mathbf{\Psi}^{(l)}), \ldots, \tau_{iM}(\mathbf{\Psi}^{(l)})$ and the i-th subject is then classified to $P_h^{(l+1)}$ when $Z_{ih}^* = 1$ for $i = 1, \ldots, n$. Because we assume that the number of components of the mixture model is known and fixed therefore in the numerical study, we designed the numerical study such that if one of the partitions becomes empty or ends up with only one subject, the SEM algorithm is stopped, and $\mathbf{\Psi}^{(l)}$ is returned.

M-Step: The $\mathbf{P}^{(l+1)}$ of the S-step is used to update the parameters of the mixture of logistic regression models in this step. First, we maximize $\mathbf{Q}_1(\pi, \mathbf{\Psi}^{(l)})$ from (4.6) subject to constraint $\sum_{j=1}^{M} \pi_j = 1$. Using the Lagrangian multiplier, it is easy to see

$$\widehat{\pi}_j^{(l+1)} = \sum_{i=1}^{n} z_{ij}^*/n = n_j/n; \qquad j = 1, \ldots, M-1, \tag{4.8}$$

where n_j denotes the number of subjects classified to $P_j{}^{(l+1)}$. To estimate the coefficients of the j-th logistic regression, we can re-write (4.7) based on partition

$\mathbf{P}^{(l+1)}$ as follows

$$\mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) = \sum_{i=1}^{n_j} \tau_{ij}(\boldsymbol{\Psi}^{(l)}) \left(y_i \mathbf{x}_i^\top \boldsymbol{\beta}_j - \log\left(1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j)\right)\right), \tag{4.9}$$

where n_j denotes the number of subjects classified to ${P_j}^{(l+1)}$. From the first derivative of (4.9) with respect to $\boldsymbol{\beta}_j$, the gradient is given by

$$\nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) = \mathbf{X}_j^\top \left(\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \boldsymbol{\beta}_j^{(l)}) \right), \tag{4.10}$$

where $\mathbf{X}_j$ and $\mathbf{y}_j$ are respectively the design matrix and vector of responses corresponding to subjects from ${P_j}^{(l+1)}$. Also, $\mathbf{g}^{-1}(\mathbf{X}_j; \boldsymbol{\beta}_j^{(l)}) = \left(g^{-1}(\mathbf{x}_1; \boldsymbol{\beta}_j^{(l)}), \ldots, g^{-1}(\mathbf{x}_{n_j}; \boldsymbol{\beta}_j^{(l)}) \right)^\top$ where $g^{-1}(.,.)$ is given by (2.9). The Hessian matrix of (4.9) is given by

$$\mathbf{H}_{\boldsymbol{\beta}_j} \left(\mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) \right) = -\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j, \tag{4.11}$$

where $\mathbf{W}_j$ is a diagonal matrix with entries

$$(w)_{ii} = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j^{(l)}) \left[1 + \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_j^{(l)}) \right]^{-2}. \tag{4.12}$$

From (4.10) and (4.11), one can use Newton-Raphson (NR) method and update $\boldsymbol{\beta}_j, j = 1, \ldots, M$ as follows

$$\boldsymbol{\beta}_j^{(l+1)} = \boldsymbol{\beta}_j^{(l)} - \mathbf{H}_{\boldsymbol{\beta}_j}^{-1} \left(\mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) \right) \nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}). \tag{4.13}$$

Lemma 4.1. *Let* $\nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)})$ *and* $\mathbf{H}_{\boldsymbol{\beta}_j} \left(\mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) \right)$ *represent the gradient and Hessian matrix of* (4.9)*. Then the iteratively re-weighted least square (IRWLS) estimate of* $\boldsymbol{\beta}_j$ *can be obtained by*

$$\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j \right)^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{V}_j,$$

where $\mathbf{W}_j$ *is diagonal weight matrix from* (4.12) *and*

$$\mathbf{V}_j = \left\{ \mathbf{X}_j \widehat{\boldsymbol{\beta}}_j^{(l)} + \mathbf{W}_j^{-1} \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)}) \right] \right\}.$$

Proof: From (4.10) and (4.11), one can obtain $\widehat{\boldsymbol{\beta}}_j^{(l+1)}$ based on partition $P_j^{(l)}$ by:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_j^{(l+1)} &= \widehat{\boldsymbol{\beta}}_j^{(l)} - \mathbf{H}_{\boldsymbol{\beta}_j}^{-1} \left(\mathbf{Q}_2(\boldsymbol{\beta}_j, \widehat{\boldsymbol{\Psi}}^{(l)}) \right) \nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2(\boldsymbol{\beta}_j, \widehat{\boldsymbol{\Psi}}^{(l)}) \\
&= \widehat{\boldsymbol{\beta}}_j^{(l)} - \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j \right)^{-1} \mathbf{X}_j^\top \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)}) \right] \\
&= \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j \right)^{-1} \mathbf{X}_j^\top \mathbf{W}_j \left\{ \mathbf{X}_j \widehat{\boldsymbol{\beta}}_j^{(l)} - \mathbf{W}_j^{-1} \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)}) \right] \right\} \\
&= \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j \right)^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{V}_j.
\end{aligned}$$

□

Finally, the IRWLS (hence referred to LS) estimate of $\boldsymbol{\Psi}$ is obtained by alternating the E-, S-, and M-steps until $|\ell(\boldsymbol{\Psi}^{(l+1)}) - \ell(\boldsymbol{\Psi}^{(l)})|$ becomes negligible.

4.3 Ridge Estimation Method

Although the LS estimation method is the most frequent method for estimating the parameters of a mixture of logistic regression models, when the covariates are linearly dependent, the LS method is seriously affected by multicollinearity. The Ridge estimation approach was proposed by (Schaefer et al., 1984) as a solution to the multicollinearity problem. We can get the Ridge estimate $\widehat{\boldsymbol{\Psi}}_R$ by maximizing the

Ridge penalized log-likelihood function of a mixture of logistic regression models. The Ridge penalized log-likelihood function is given by

$$\ell^R(\beta) = \ell(\beta) - \frac{1}{2}k\boldsymbol{\beta}^\top\boldsymbol{\beta}, \tag{4.14}$$

where $\ell(\beta)$ is the incomplete log-likelihood function (2.10) and k is the Ridge parameter. There is no closed form for $\widehat{\boldsymbol{\Psi}}_R$ using (4.14) similar to Subsection 4.2.1. Accordingly, we introduce the latent variables $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_M)$ and run a SEM algorithm on the entire data $(\mathbf{X}, \mathbf{y}, \mathbf{Z})$ to get $\widehat{\boldsymbol{\Psi}}_R$. To do so, we use Subsection 4.2.1 to implement the E- and S- steps of the Ridge estimation approach.

In the M-step, the mixing proportion $\widehat{\pi}_j, j = 1, \ldots, M$ can be obtained from (4.8). To obtain the estimate of logistic coefficients, we maximize the conditional expectation log-likelihood subject to the ridge penalty as follows

$$\mathbf{Q}_2^R(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) = \mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) - k_j\boldsymbol{\beta}_j^\top\boldsymbol{\beta}_j/2, \tag{4.15}$$

where $\mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)})$ comes from (4.9) and λ_j is the Ridge parameter in j-th component of the mixture.

Lemma 4.2. *Under the assumptions of Lemma 4.1. The Ridge estimator* $\widehat{\boldsymbol{\beta}}_R^{(l+1)} = (\widehat{\boldsymbol{\beta}}_{R,1}^{(l+1)}, \ldots, \widehat{\boldsymbol{\beta}}_{R,M}^{(l+1)})$ *using the IRWLS method is updated by*

$$\widehat{\boldsymbol{\beta}}_{R,j} = \left(\mathbf{X}_j^\top\mathbf{W}_j\mathbf{X}_j + k_j\mathbb{I}\right)^{-1}\mathbf{X}_j^\top\mathbf{W}_j\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}}_{LS,j},$$

where $\widehat{\boldsymbol{\beta}}_{LS,j}$ *is given by Lemma 4.1.*

Proof: Taking the first and second derivative from (4.15) wrt $\boldsymbol{\beta}_j$, the Ridge gradient and Ridge Hessian matrix are given by

$$\nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2^R(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) = \mathbf{X}_j^\top \left(\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \boldsymbol{\beta}_j^{(l)}) \right) - k_j \boldsymbol{\beta}_j, \tag{4.16}$$

$$\mathbf{H}_{\boldsymbol{\beta}_j} \left(\mathbf{Q}_2^R(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) \right) = -\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j - k_j \mathbb{I}. \tag{4.17}$$

Let $\mathbf{U}_j = \mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j + k_j \mathbb{I}$. From (4.16) and (4.17), the Ridge estimate $\widehat{\boldsymbol{\beta}}_{R,j}^{(l+1)}$ can be updated by an iteratively re-weighted least squares as follows

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_j^{(l+1)} &= \widehat{\boldsymbol{\beta}}_j^{(l)} - \mathbf{H}_j^{-1} \left(\mathbf{Q}_2^R(_j, \widehat{\boldsymbol{\Psi}}^{(l)}) \right) \nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2^R(_j, \widehat{\boldsymbol{\Psi}}^{(l)}) \\
&= \widehat{\boldsymbol{\beta}}_j^{(l)} + \mathbf{U}_j^{-1} \left\{ \mathbf{X}_j^\top \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)}) \right] - k_j \widehat{\boldsymbol{\beta}}_j^{(l)} \right\} \\
&= \mathbf{U}_j^{-1} \mathbf{U}_j \widehat{\boldsymbol{\beta}}_j^{(l)} - k_j \mathbf{U}_j^{-1} \widehat{\boldsymbol{\beta}}_j^{(l)} + \mathbf{U}_j^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{W}_j^{-1} \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)}) \right] \\
&= \mathbf{U}_j^{-1} \mathbf{X}_j^\top \mathbf{W}_j \left\{ \mathbf{X}_j \widehat{\boldsymbol{\beta}}_j^{(l)} + \mathbf{W}_j^{-1} \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)}) \right] \right\} \\
&= \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j + k_j \mathbb{I} \right)^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{V}_j.
\end{aligned}$$

□

We estimate the Ridge parameter k_j by $\widehat{k}_j = (p+1)/\widehat{\boldsymbol{\beta}}_{LS,j}^\top \widehat{\boldsymbol{\beta}}_{LS,j}$ using (Inan and Erdogan, 2013) where p represents the number of explanatory variables and $\widehat{\boldsymbol{\beta}}_{LS,j}$ is the LS estimate of $\boldsymbol{\beta}_j$. Finally, the estimate of $\widehat{\boldsymbol{\Psi}}_R$ is obtained by alternating the E-, S-, and M-steps until $|\ell(\boldsymbol{\Psi}_R^{(l+1)}) - \ell(\boldsymbol{\Psi}_R^{(l)})|$ becomes negligible.

4.4 Liu-type Estimation Method

The Ridge approach may not adequately handle the severe ill-conditioned design matrix when multicollinearity is present. The Liu-type (LT) approach was proposed by Liu (Liu, 2003) and Inan et al (Inan and Erdogan, 2013) as a solution to the problem in regression and logistic regression, respectively. In the presence of multicollinearity, we offer the LT technique for estimating the parameters of a mixture of logistic regression models. To do that, we replace the Ridge penalty $0 = k^{1/2}\boldsymbol{\beta} + \epsilon'$ by the LT penalty

$$(-\frac{d}{k^{1/2}})\widehat{\boldsymbol{\beta}} = k^{1/2}\boldsymbol{\beta} + \epsilon', \tag{4.18}$$

where $\widehat{\boldsymbol{\beta}}$ can be any estimator of coefficients and $d \in \mathbb{R}$ and $k > 0$ are two parameters of the LT estimation method. Throughout this chapter, we use $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_R$ in LT penalty (4.18). We see $(\boldsymbol{X}, \boldsymbol{y})$ as incomplete data and convert them into complete data $(\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{Z})$, where $\boldsymbol{Z}$ includes the missing component memberships, similar to the LS approach (explained in Subsection 4.2.1). The LT estimate of the parameters of the mixture of logistic regression models is then determined using the SEM technique. Here, the E- and S-steps are treated similarly to the LS and Ridge estimation methods.

In the M-step, we start with the classified data from the S-step and use (4.8) to estimate the mixing proportions. Later, we maximize $\mathbf{Q}_2(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)})$ subject to LT penalty (4.18) to estimate the coefficients within each partition $P_j^{(l+1)}$ for $j = 1, \ldots, M$.

Lemma 4.3. *Under the assumptions of Lemma 4.1. The LT estimator* $\widehat{\boldsymbol{\beta}}_{LT}^{(l+1)} = (\widehat{\boldsymbol{\beta}}_{LT,1}^{(l+1)}, \ldots, \widehat{\boldsymbol{\beta}}_{LT,M}^{(l+1)})$ *using the IRWLS method is updated by*

$$\widehat{\boldsymbol{\beta}}_{LT,j} = \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j + k_j \mathbb{I}\right)^{-1} \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{V}_j - d_j \widehat{\boldsymbol{\beta}}_{R,j}\right),$$

where $\mathbf{W}_j$ *and* $\mathbf{V}_j$ *are given by Lemma 4.1 and* $\widehat{\boldsymbol{\beta}}_{R,j}$ *is calculated from Lemma 4.2.*

Proof: It is easy to show that the gradient and Ridge Hessian matrix under the LT estimation method are given by

$$\nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2^{LT}(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)}) = \mathbf{X}_j^\top \left(\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \boldsymbol{\beta}_j^{(l)})\right) - d_j \widehat{\boldsymbol{\beta}}_{R,j} - k_j \boldsymbol{\beta}_j, \tag{4.19}$$

$$\mathbf{H}_{\boldsymbol{\beta}_j} \left(\mathbf{Q}_2^{LT}(\boldsymbol{\beta}_j, \boldsymbol{\Psi}^{(l)})\right) = -\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j - k_j \mathbb{I}. \tag{4.20}$$

Let $\mathbf{U}_j = \mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j + k_j \mathbb{I}$. From (4.19) and (4.20), the LT estimate $\widehat{\boldsymbol{\beta}}_{LT,j}^{(l+1)}$ can be updated by an iteratively re-weighted least squares as follows

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_j^{(l+1)} &= \widehat{\boldsymbol{\beta}}_j^{(l)} - \mathbf{H}_{\boldsymbol{\beta}_j}^{-1} \left(\mathbf{Q}_2^{LT}(\boldsymbol{\beta}_j, \widehat{\boldsymbol{\Psi}}^{(l)})\right) \nabla_{\boldsymbol{\beta}_j} \mathbf{Q}_2^{LT}(\boldsymbol{\beta}_j, \widehat{\boldsymbol{\Psi}}^{(l)}) \\
&= \widehat{\boldsymbol{\beta}}_j^{(l)} + \mathbf{U}_j^{-1} \left\{\mathbf{X}_j^\top \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)})\right] - k_j \widehat{\boldsymbol{\beta}}_j^{(l)} - d_j \widehat{\boldsymbol{\beta}}_{R,j}\right\} \\
&= \mathbf{U}_j^{-1} \mathbf{U}_j \widehat{\boldsymbol{\beta}}_j^{(l)} - k_j \mathbf{U}_j^{-1} \widehat{\boldsymbol{\beta}}_j^{(l)} + \mathbf{U}_j^{-1} \mathbf{X}_j^\top \mathbf{W}_j \mathbf{W}_j^{-1} \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)})\right] - d_j \mathbf{U}_j^{-1} \widehat{\boldsymbol{\beta}}_{R,j} \\
&= \mathbf{U}_j^{-1} \mathbf{X}_j^\top \mathbf{W}_j \left\{\mathbf{X}_j \widehat{\boldsymbol{\beta}}_j^{(l)} + \mathbf{W}_j^{-1} \left[\mathbf{y}_j - \mathbf{g}^{-1}(\mathbf{X}_j; \widehat{\boldsymbol{\beta}}_j^{(l)})\right]\right\} - d_j \mathbf{U}_j^{-1} \widehat{\boldsymbol{\beta}}_{R,j} \\
&= \left(\mathbf{X}_j^\top \mathbf{W}_j \mathbf{X}_j + k_j \mathbb{I}\right)^{-1} \left\{\mathbf{X}_j^\top \mathbf{W}_j \mathbf{V}_j - d_j \widehat{\boldsymbol{\beta}}_{R,j}\right\}.
\end{aligned}$$

□

There are several techniques to estimate k_j, according to Schaefer et al. (1984) and Inan and Erdogan (2013). Here, we use $\widehat{k}_j = (p+1)/\widehat{\boldsymbol{\beta}}_{R,j}^{\top}\widehat{\boldsymbol{\beta}}_{R,j}$ to estimate the parameters, where p is the number of explanatory variables and $\widehat{\boldsymbol{\beta}}_{R,j}$ is the Ridge estimate of $\boldsymbol{\beta}_j$. We employ the operational technique of Inan and Erdogan (2013) to estimate the bias correction parameters d_j by maximizing the mean square errors (MSE) of $\widehat{\boldsymbol{\beta}}_{LT,j}$ inside each partition $P_j^{(l+1)}$ once k_j has been estimated. It is easy to show that.

$$\text{MSE}(\widehat{\boldsymbol{\beta}}_{LT,j}) = tr\left[\text{Var}(\widehat{\boldsymbol{\beta}}_{LT,j})\right] + ||\mathbb{E}(\widehat{\boldsymbol{\beta}}_{LT,j}) - \boldsymbol{\beta}_j||_2^2,$$

where

$$tr\left[\text{Var}(\widehat{\boldsymbol{\beta}}_{LT,j})\right] = tr\Big[\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{X}_j + k_j\mathbb{I}\right)^{-1}\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{V}_j - d_j\mathbb{I}\right)\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{X}_j + k_j\mathbb{I}\right)^{-1}$$
$$\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{X}_j\right)\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{X}_j + k_j\mathbb{I}\right)^{-1}$$
$$\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{V}_j - d_j\mathbb{I}\right)\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{X}_j + k_j\mathbb{I}\right)^{-1}\Big],$$

and

$$||\mathbb{E}(\widehat{\boldsymbol{\beta}}_{LT,j}) - \boldsymbol{\beta}_j||_2^2 = ||\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{X}_j + k_j\mathbb{I}\right)^{-1}\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{V}_j - d_j\mathbb{I}\right)\left(\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{X}_j + k_j\mathbb{I}\right)^{-1}$$
$$\mathbf{X}_j^{\top}\mathbf{W}_j\mathbf{g}^{-1}(\mathbf{X}_j;\boldsymbol{\beta}_j) - \boldsymbol{\beta}_j||_2^2.$$

As you can see, the true value of the parameters $\boldsymbol{\beta}_j$ affects $\text{MSE}(\widehat{\boldsymbol{\beta}}_{LT,j})$. As a result, while calculating the bias correction parameters of the LT method $\boldsymbol{d} = (d_1, d_2, \ldots, d_M)$, the true $\boldsymbol{\beta}_j$ are replaced with $\widehat{\boldsymbol{\beta}}_{R,j}$ and according to Inan and Erdogan (2013), we choose the optimum tuning parameter d value which minimizes

the MSE($\widehat{\boldsymbol{\beta}}_{LT,j}$). Finally, the E-, S- and M-steps of SEM algorithm under LT method is alternated until $|\ell(\boldsymbol{\Psi}_{LT}^{(l+1)}) - \ell(\boldsymbol{\Psi}_{LT}^{(l)})|$ becomes negligible.

Chapter 5

Numerical Studies

This chapter presents two different numerical studies to compare the performance of the ML, Ridge, and LT methods in estimating the parameters in both a mixture of logistic regression models and a mixture of regression models in multicollinearity. Finally, we applied our proposed methods to analyze the bone disorder status of women aged 50 and older.

This chapter is organized as follows. Sections 5.1.1 and 5.1.2 assess the performance of the estimation methods via two different simulation studies for mixture of logistic regression. Sections 5.2.1 and 5.2.2 assess the performance of the estimation methods via two different simulation studies for mixture of regression models. Section 5.3 describes the real data example for mixture of logistic regression and the mixture of regression models

5.1 Simulation Studies for Logistic Regression

This section compares the performance of the ML, Ridge, and LT approaches in estimating the parameters of a mixture of logistic regression models in the presence

of multicollinearity using two simulation studies. We look at how the sample size, multicollinearity level, and several components in the mixture of logistic regressions affect the proposed estimate approaches. We start by assuming that the underlying population is made up of two logistic regression models. The approaches' performance is then investigated in the second simulation, where the population consists of three logistic regression components.

5.1.1 Simulation Study 1

In the first simulation study, we employed two parameters ϕ and ρ, to induce multicollinearity in the mixture of logistic regressions, as described by Inan and Erdogan (2013). We also took into account the component logistic regressions, which include four covariates $(\mathbf{x}_1, \ldots, \mathbf{x}_4)$, where ϕ and ρ denote the levels of correlation between the first and last two predictors in the mixture model. We first generated random numbers $\{w_{ij}, i = 1, \ldots, n; j = 1, \ldots, 5\}$ from the standard normal distribution and then simulated the covariates as follows

$$x_{i,j_1} = (1 - \phi^2)w_{i,j_1} + \phi w_{i,5}, \quad j_1 = 1, 2,$$

$$x_{i,j_2} = (1 - \rho^2)w_{i,j_2} + \rho w_{i,5}, \quad j_2 = 3, 4,$$

where we used $\phi = \{0.85, 0.95, 0.98\}$ and $\rho = \{0.9, 0.95, 0.99\}$ to simulate the multicollinearity in the mixture of logistic regressions. We then generated the binary responses from logistic regression $p_1(\mathbf{x}_i; \boldsymbol{\beta}_{01})$ with probability $\pi_0 = 0.7$ and from logistic regression $p_2(\mathbf{x}_i; \boldsymbol{\beta}_{02})$ with probability 0.3 where $p_j(\cdot; \cdot)$ is given by (4.2) and $\psi_0 = (\boldsymbol{\pi_0}, \boldsymbol{\beta}_0)$ with $\pi_0 = (0.7, 0.3)$, $\boldsymbol{\beta_0} = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02})$ where, $\boldsymbol{\beta}_{01} = (1, 3, 4, 5, 6)$ and $\boldsymbol{\beta}_{02} = (-1, -1, -2, -3, -5)$. These shrinkage methods required two starting

values. Note that we must assign a value for the initial starting point for each of the ML, Ridge, or Liu-type methods. However, since our main goal is to compare the performance of these algorithms, we choose the same initial values, close to the true parameters, for all the estimation methods, including ML, Ridge, and Liu-type methods in all the simulation studies. The initial values that we used in this simulation are $\boldsymbol{\beta}_{01(initial)} = \boldsymbol{\beta}_0 + \mathbf{2}$, $\boldsymbol{\beta}_{02(initial)} = \boldsymbol{\beta}_0 - \mathbf{2}$, $\pi_{01} = (0.5, 0.5)$ and $\pi_{02} = (0.3, 0.7)$. The initial values are considered fixed for all methods (ML, Ridge and Liu) for fair comparison.

To investigate the estimation performance of $(\widehat{\pi}, \widehat{\boldsymbol{\beta}})$, we used the sum of squared errors (SSE) of the estimates and measured $\sqrt{\text{SSE}(\widehat{\boldsymbol{\beta}})} = \left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right]^{1/2}$ and $\sqrt{\text{SSE}(\widehat{\pi})} = [(\widehat{\pi} - \widehat{\pi}_0)^2]^{1/2}$ where $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)^\top$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02})^\top$. We first estimated the mixture model parameters using a training sample of size n to examine the classification performance of the approaches. From the underlying mixture of two logistic regression models, we constructed a validation set of size 100 (independent of the training data). The trained model was then used to predict the binary response of the validation set. We computed the prediction measures of Error = (FP+FN)/(TP+TN+FP+FN), Sensitivity = (TP)/(TP+FN) and Specificity = (TN)/(TN+FP) where FP, FN, TP and TN stand for false positive, false negative, true positive and true negative entries in the confusion matrix, respectively. In order to study the effect of sample size, we investigate the estimation and prediction performance with different sample sizes, including $n = \{25, 40, 100\}$. Note that when the sample size is small, say $n = 25$, even the convergence rate of LS is reduced. Furthermore, it is well-known that the convergence rate of SEM

is not guaranteed regardless of the size of n (Faria and Soromenho, 2010). In the presence of multicollinearity, the estimation methods may result in skewed and outlier estimation and classification results. The results also change dramatically from one replication to another. For these reasons, to simulate the performance of the proposed methods, we first generated data from the underlying mixture models. We then estimated the population parameters and computed the $\sqrt{\text{SSE}}$s of the estimates and the classification measures.

We note that our objective is to compare the performance of the ML, Ridge, and LT over a fixed sample size. However, we did not intend to compare the performance of a method when the sample changes; in this case, it would be preferred to use $\sqrt{\text{MSE}}$ instead of $\sqrt{\text{SSE}}$.

To better compare the asymmetry in the proposed estimations and classification methods, we replicated the entire data generation, estimation, and classification procedures 2000 times using the ML, Ridge and LT methods. Then, we computed the 2.5%, 50% and 97.5% percentiles of the $\sqrt{\text{SSE}}$s and evaluated the Error, Sensitivity and Specificity for the classification. We presented the lower (L) and upper (U) bounds of the estimation and classification intervals by 2.5 and 97.5 percentiles of the corresponding criterion, respectively. From Tables 5.1 and 5.7, almost similar results are observed. When the sample size is large, we see the performance of LT and Ridge is similar to LT shrinkage, however, as the sample size decrease, the LT shrinkage method appears more reliable than LS and Ridge.

Table 5.1: The median (M), lower (L) and upper (U) bounds of 95% intervals for $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the methods in estimation and prediction of the mixture of two logistic regressions when $n = 25$ and $\rho = 0.9$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	223	46	1×10^6	.46	.28	.68	.55	.20	.83	.55	.19	.83
		π	.14	.02	.58									
	Ridge	β	32	19	203	.44	.28	.66	.55	.19	.84	.57	.23	.86
		π	.22	.02	.70									
	LT	β	30	21	36	.46	.30	.60	.56	.29	.81	.55	.26	.78
		π	.3	.02	.70									
0.95	ML	β	529	66	1×10^6	.46	.28	.68	.55	.21	.83	.56	.21	.84
		π	.14	.02	.58									
	Ridge	β	32	18	196	.44	.28	.66	.55	.19	.83	.56	.22	.85
		π	.22	.02	.70									
	LT	β	30	21	36	.46	.30	.60	.55	.30	.81	.54	.27	.79
		π	.22	.02	.70									
0.98	ML	β	920	96	2×10^6	.44	.28	.68	.54	.22	.82	.56	.19	.86
		π	.14	.02	.58									
	Ridge	β	31	19	207	.44	.26	.68	.56	.19	.83	.56	.23	.85
		π	.26	.02	.70									
	LT	β	30	21	35	.46	.30	.62	.56	.28	.81	.54	.28	.80
		π	.24	.02	.70									

Tables 5.1-5.9 show the results of the simulation study. The ML approach performs somewhat better than the Ridge and LT methods in estimating mixing proportions. This is based on the fact that the Ridge and LT approaches are biassed shrinkage methods. These shrinkage methods are aimed to overcome multicollinearity and improve the analysis of the model's coefficients by integrating a bias into the estimation. While the multicollinearity had a major impact on the ML estimates, the Ridge and LT estimates looked to be far more reliable in determining the mixture model coefficients. We also observe that $\widehat{\boldsymbol{\beta}}_{LT}$ significantly outperforms $\widehat{\boldsymbol{\beta}}_{R}$ where the intervals for the $\sqrt{\text{SSE}}$ of $\widehat{\boldsymbol{\beta}}_{R}$ account for 5-10 times wider than those of $\widehat{\boldsymbol{\beta}}_{LT}$. Similar to the findings of (Inan and Erdogan, 2013), the classification performances of Error, Specificity (SP), and Sensitivity (SN) under the three methods are almost the same. We see that the SP and SN are a little low when we apply them

to the three methods; because of that, the simulation study structure is based on high multicollinearity, which is why all the methods suffer from high multicollinearity. Inan and Erdogan (2013) reported only the mean of classification measures for one logistic in the presence of high multicollinearity. Because the results of the estimation and classification are highly skewed, unlike Inan and Erdogan (2013), we reported the median and 95% intervals for $\sqrt{\text{SSE}}$, Error, Sensitivity and Specificity to better investigate the performance of the estimators. Interestingly, if the sample size is small ($n = 25$), the LT shrinkage method appears more reliable relative to LS and Ridge in estimating the coefficients of all the logistic components. However, if the sample size is large enough ($n = 100$), LS and Ridge could perform as well as LT shrinkage.

Table 5.2: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 25$ and $\rho = 0.95$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	470	62	1×10^6	.44	.28	.68	.54	.21	.83	.56	.20	.84
		π	.14	.02	.58									
	Ridge	β	32	20	233	.44	.28	.68	.56	.19	.84	.57	.23	.86
		π	.22	.02	.70									
	LT	β	30	21	36	.46	.30	.60	.55	.28	.81	.55	.27	.78
		π	.3	.02	.70									
0.95	ML	β	705	102	1×10^6	.46	.28	.68	.55	.22	.83	.55	.20	.85
		π	.14	.02	.54									
	Ridge	β	31	19	228	.44	.28	.68	.55	.17	.83	.57	.23	.86
		π	.22	.02	.70									
	LT	β	30	21	36	.46	.30	.60	.56	.29	.81	.55	.27	.79
		π	.3	.02	.70									
0.98	ML	β	1143	120	2×10^6	.46	.28	.68	.55	.21	.83	.55	.19	.83
		π	.14	.02	.58									
	Ridge	β	31	18	168	.44	.26	.68	.56	.19	.83	.56	.18	.86
		π	.22	.02	.70									
	LT	β	30	21	37	.46	.30	.62	.56	.29	.81	.54	.26	.78
		π	.3	.02	.70									

Table 5.3: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 25$ and $\rho = 0.99$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	1383	118	1×10^6	.46	.28	.68	.54	.21	.83	.56	.20	.85
		π	.14	.02	.58									
	Ridge	β	32	19	181	.44	.28	.68	.56	.19	.84	.56	.23	.86
		π	.22	.02	.70									
	LT	β	30	21	35	.46	.30	.60	.56	.29	.80	.55	.27	.79
		π	.3	.02	.70									
0.95	ML	β	1651	146	1×10^6	.46	.28	.68	.55	.21	.83	.55	.20	.86
		π	.14	.02	.54									
	Ridge	β	31	19	171	.44	.28	.68	.56	.21	.84	.56	.21	.86
		π	.22	.02	.70									
	LT	β	30	21	37	.46	.30	.62	.55	.29	.80	.55	.29	.79
		π	.3	.02	.70									
0.98	ML	β	2387	248	3×10^6	.44	.28	.68	.55	.21	.83	.55	.17	.86
		π	.14	.02	.58									
	Ridge	β	31	18	326	.44	.26	.68	.56	.19	.83	.57	.22	.85
		π	.22	.02	.70									
	LT	β	30	21	37	.46	.30	.60	.56	.29	.80	.54	.29	.79
		π	.3	.02	.70									

Table 5.4: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 40$ and $\rho = 0.9$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	270	47	2×10^5	.46	.28	.68	.52	.20	.82	.56	.19	.85
		π	.15	.00	.60									
	Ridge	β	32	22	230	.44	.28	.66	.56	.20	.83	.56	.25	.82
		π	.25	.025	.70									
	LT	β	30	22	37	.44	.30	.60	.56	.31	.79	.55	.29	.79
		π	.3	.025	.70									
0.95	ML	β	434	73	3×10^5	.46	.28	.68	.54	.19	.82	.55	.20	.85
		π	.15	.00	.60									
	Ridge	β	32	22	239	.44	.28	.68	.55	.21	.81	.56	.23	.83
		π	.25	.025	.70									
	LT	β	30	22	36	.44	.30	.60	.57	.31	.79	.55	.30	.78
		π	.25	.025	.70									
0.98	ML	β	740	112	4×10^5	.46	.28	.70	.53	.20	.83	.54	.19	.85
		π	.15	.00	.60									
	Ridge	β	32	21	222	.44	.28	.68	.57	.21	.82	.55	.24	.83
		π	.25	.025	.70									
	LT	β	30	22	37	.44	.30	.60	.56	.31	.81	.55	.30	.79
		π	.25	.025	.70									

Table 5.5: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 40$ and $\rho = 0.95$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	380	60	2×10^5	.46	.28	.70	.52	.20	.80	.54	.19	.85
		π	.125	.00	.57									
	Ridge	β	32	23	263	.44	.28	.68	.56	.24	.82	.55	.22	.84
		π	.25	.025	.70									
	LT	β	30	22	36	.44	.30	.60	.56	.31	.80	.56	.32	.79
		π	.25	.05	.70									
0.95	ML	β	628	99	4×10^5	.48	.28	.68	.52	.20	.83	.54	.18	.83
		π	.15	.00	.60									
	Ridge	β	32	22	172	.44	.28	.68	.56	.21	.82	.55	.22	.85
		π	.25	.025	.70									
	LT	β	30	21	37	.44	.30	.62	.57	.30	.79	.55	.30	.79
		π	.25	.025	.70									
0.98	ML	β	942	126	5×10^5	.46	.28	.70	.52	.21	.83	.55	.17	.85
		π	.125	.00	.60									
	Ridge	β	31	21	292	.44	.28	.68	.55	.21	.82	.56	.22	.84
		π	.25	.025	.70									
	LT	β	30	22	40	.44	.30	.60	.56	.30	.79	.55	.30	.78
		π	.275	.025	.70									

Table 5.6: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 40$ and $\rho = 0.99$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	1132	142	4×10^5	.48	.28	.70	.52	.19	.82	.55	.18	.85
		π	.15	.00	.6									
	Ridge	β	32	22	280	.44	.28	.68	.56	.17	.83	.56	.23	.84
		π	.25	.025	.70									
	LT	β	30	22	39	.44	.30	.60	.55	.30	.79	.56	.30	.78
		π	.25	.05	.70									
0.95	ML	β	1333	195	6×10^5	.48	.28	.70	.52	.18	.81	.54	.18	.84
		π	.15	.00	.57									
	Ridge	β	32	21	282	.44	.28	.68	.56	.19	.83	.56	.23	.85
		π	.25	.025	.70									
	LT	β	30	22	38	.44	.30	.60	.56	.31	.79	.56	.30	.77
		π	.25	.025	.70									
0.98	ML	β	1900	267	5×10^5	.46	.28	.68	.52	.20	.80	.54	.19	.86
		π	.15	.00	.57									
	Ridge	β	31	20	265	.44	.28	.70	.56	.22	.82	.56	.22	.83
		π	.25	.025	.70									
	LT	β	30	22	42	.44	.30	.60	.56	.31	.79	.55	.30	.78
		π	.275	.025	.70									

Table 5.7: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 100$ and $\rho = 0.9$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	186	43	1×10^4	.46	.28	.70	.54	.19	.81	.55	.21	.83
		π	.12	.01	.65									
	Ridge	β	30	20	52	.44	.30	.60	.57	.33	.80	.56	.30	.78
		π	.18	.01	.67									
	LT	β	30	22	36	.44	.30	.58	.56	.35	.77	.55	.32	.76
		π	.26	.01	.70									
0.95	ML	β	291	66	2×10^4	.46	.28	.68	.54	.19	.81	.56	.21	.85
		π	.13	.00	.66									
	Ridge	β	30	20	54	.44	.30	.62	.56	.30	.78	.56	.30	.78
		π	.21	.01	.68									
	LT	β	29	22	35	.44	.30	.60	.57	.33	.78	.56	.33	.76
		π	.25	.01	.70									
0.98	ML	β	511	101	3×10^5	.46	.28	.68	.54	.17	.82	.56	.21	.85
		π	.12	.00	.65									
	Ridge	β	30	19	52	.44	.28	.61	.57	.32	.79	.56	.30	.79
		π	.18	.01	.68									
	LT	β	29	22	36	.44	.30	.58	.57	.35	.77	.56	.33	.78
		π	.24	.01	.70									

Table 5.8: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 100$ and $\rho = 0.95$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	263	57	2×10^4	.46	.28	.68	.54	.19	.82	.55	.21	.83
		π	.12	.00	.64									
	Ridge	β	30	20	53	.44	.30	.62	.57	.32	.77	.56	.30	.79
		π	.20	.01	.67									
	LT	β	30	22	36	.44	.30	.58	.57	.34	.79	.56	.33	.77
		π	.26	.01	.70									
0.95	ML	β	393	76	2×10^4	.46	.28	.70	.53	.18	.81	.54	.19	.85
		π	.12	.01	.64									
	Ridge	β	30	19	50	.44	.30	.62	.57	.32	.80	.57	.30	.79
		π	.22	.01	.70									
	LT	β	29	22	35	.44	.30	.58	.57	.33	.77	.55	.33	.76
		π	.25	.01	.68									
0.98	ML	β	623	121	4×10^4	.46	.28	.68	.54	.22	.81	.54	.19	.83
		π	.12	.00	.64									
	Ridge	β	30	18	48	.44	.30	.62	.57	.30	.78	.56	.30	.78
		π	.22	.01	.70									
	LT	β	29	21	36	.44	.30	.58	.57	.35	.77	.56	.33	.77
		π	.24	.01	.70									

Table 5.9: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of two logistic regressions when $n = 100$ and $\rho = 0.99$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	800	118	5×10^4	.46	.28	.68	.54	.20	.81	.55	.20	.85
		π	.12	.00	.64									
	Ridge	β	30	19	59	.44	.28	.62	.56	.30	.78	.56	.30	.80
		π	.19	.01	.67									
	LT	β	29	22	37	.44	.30	.58	.57	.33	.77	.56	.32	.77
		π	.25	.01	.68									
0.95	ML	β	951	176	6×10^4	.46	.28	.70	.54	.19	.81	.56	.19	.85
		π	.12	.00	.64									
	Ridge	β	30	18	50	.44	.28	.62	.57	.31	.80	.56	.29	.78
		π	.23	.01	.68									
	LT	β	29	21	36	.44	.30	.58	.57	.35	.78	.56	.33	.76
		π	.24	.01	.68									
0.98	ML	β	1331	268	8×10^4	.46	.28	.70	.54	.21	.81	.56	.20	.84
		π	.12	.00	.61									
	Ridge	β	30	17	80	.44	.30	.64	.57	.27	.79	.56	.29	.79
		π	.20	.01	.68									
	LT	β	29	20	37	.44	.30	.58	.57	.33	.77	.56	.33	.77
		π	.24	.01	.68									

5.1.2 Simulation Study 2

The second simulation looks at how well the estimation approaches perform when the population is made up of three logistic regression models with two covariates. Assuming the correlation level $\phi = \{0.85, 0.95, 0.99\}$, we generated the covariates and binary responses as described above from the mixture population when $\pi_0 = (0.3, 0.4, 0.3)$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \boldsymbol{\beta}_{03})$ with $\boldsymbol{\beta}_{01} = (2.85, -10, -5.11)$, $\boldsymbol{\beta}_{02} = (10, 9.90, 5.11)$ and $\boldsymbol{\beta}_{03} = (-3.84, 9.90, 5.11)$. The initial values that we used in this simulation are $\boldsymbol{\beta}_{01(initial)} = \boldsymbol{\beta}_0 + \mathbf{2}$, $\boldsymbol{\beta}_{02(initial)} = \boldsymbol{\beta}_0 + \mathbf{1}$, $\pi_{01} = (0.3, 0.4, 0.3)$ and $\pi_{02} = (0.2, 0.4, 0.4)$ and all are fixed for all methods (ML, Ridge and Liu). Similar to the setting of the first study, we computed the medians and 95% in-

tervals for the estimation and classification measures using different sample sizes, $n = \{50, 100\}$. The results of this study is presented in Tables 5.10 and 5.11. While the three methods' prediction performance is nearly identical, the Ridge and LT methods produced more reliable estimates for the mixture of logistic regression model's coefficients. As you can see ML estimates become extremely unreliable in estimating the parameters, for example from Table 5.10 the 95% interval of $\sqrt{\text{SSE}}$ when $\phi = 0.85$ is between $[32, 6 \times 10^5]$. Unlike, the Ridge and LT performs more reliably in estimating parameters, for example the 95% interval of $\sqrt{\text{SSE}}$ is between [44,69] while LT is between [46,65].

In addition, when estimating the coefficients of the mixture model, the LT estimates almost consistently outperform their Ridge counterparts. As a result, we developed a method such that we got a good prediction performance, and we have a proposal that we solved the estimation problem too.

Table 5.10: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of three logistic regressions when $n = 50$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	139	32	6×10^5	.45	.32	.65	.60	.18	.84	.48	.20	.77
		π	.34	.06	.70									
	Ridge	β	59	44	69	.43	.31	.56	.68	.37	.90	.43	.12	.70
		π	.44	.14	.80									
	LT	β	60	46	65	.42	.30	.55	.69	.38	.93	.42	.12	.72
		π	.42	.16	.74									
0.95	ML	β	169	32	7×10^5	.45	.32	.64	.60	.18	.84	.49	.20	.80
		π	.32	.06	.70									
	Ridge	β	58	43	68	.43	.31	.55	.67	.38	.91	.44	.15	.72
		π	.44	.13	.80									
	LT	β	60	45	65	.42	.30	.55	.70	.37	.95	.43	.11	.72
		π	.42	.16	.72									
0.99	ML	β	255	38	2×10^5	.45	.32	.64	.61	.20	.84	.49	.22	.79
		π	.34	.06	.70									
	Ridge	β	58	43	69	.42	.30	.56	.68	.36	.92	.43	.14	.72
		π	.42	.14	.67									
	LT	β	60	45	65	.42	.29	.55	.70	.34	.95	.43	.09	.76
		π	.42	.16	.74									

Table 5.11: The median (M) and 95% intervals for the $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the ML, Ridge and LT methods in estimation and prediction of the mixture of three logistic regressions when $n = 100$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
ϕ	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
0.85	ML	β	115	34	1×10^4	.44	.32	.63	.63	.20	.87	.47	.19	.77
		π	.38	.07	.82									
	Ridge	β	58	44	68	.42	.31	.54	.68	.42	.91	.43	.15	.70
		π	.47	.16	.84									
	LT	β	60	42	66	.41	.30	.55	.70	.40	.93	.43	.13	.72
		π	.45	.18	.79									
0.95	ML	β	131	37	1×10^5	.44	.31	.63	.64	.20	.87	.47	.18	.76
		π	.39	.07	.8									
	Ridge	β	58	43	69	.41	.30	.53	.68	.44	.92	.44	.17	.68
		π	.48	.16	.84									
	LT	β	60	43	66	.41	.29	.53	.71	.42	.94	.43	.13	.71
		π	.46	.18	.79									
0.99	ML	β	201	47	2×10^4	.43	.31	.62	.64	.22	.88	.47	.18	.78
		π	.38	.08	.81									
	Ridge	β	58	43	68	.41	.30	.54	.68	.43	.92	.45	.16	.69
		π	.47	.17	.85									
	LT	β	60	45	65	.41	.29	.53	.71	.40	.95	.44	.13	.74
		π	.47	.18	.79									

5.2 Simulation Studies for Linear of Regression Models

This section examines the performance of the ML, Ridge, and LT techniques in estimating the parameters of a mixture of regression models in the presence of multicollinearity by using two simulated studies. We investigate how the sample size, multicollinearity level, and multiple components in the mixture of regressions affect the proposed estimate methods. To begin, we assume the underlying population consists of two regression models. The performance of the techniques is then examined in a second simulation, in which the population is made up of three regression components.

5.2.1 Simulation Study 1

In the first simulation research, we employed one parameter ρ, to induce multicollinearity in the mixture of regressions, as described by Inan and Erdogan (2013). The component regressions include four covariates $(\mathbf{x}_1, \ldots, \mathbf{x}_4)$, where ρ denote the level of correlation between the four predictors in the mixture model. We first generated random numbers $\{w_{ij}, i = 1, \ldots, n; j = 1, \ldots, 4\}$ from the standard normal distribution and then simulated the covariates as follows

$$x_{i,j} = (1 - \rho^2)w_{i,j} + \rho w_{i,5}, \quad j = 1, 2, 3, 4,$$

where we used $\rho = \{0.88, 0.9, 0.95, 0.97, 0.99\}$ to simulate the multicollinearity in the mixture of regressions. We then generated the responses form regression (3.1) with probability $\pi_0 = (0.7, 0.3)$ and $\Psi_0 = (\pi_0, \boldsymbol{\beta}_0, \boldsymbol{\sigma}_0^2)$ where, $\boldsymbol{\beta_0} = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02})$ such

that, $\boldsymbol{\beta}_{01} = (1, 3, 4, 5, 6)$, $\boldsymbol{\beta}_{02} = (-1, -1, -2, -3, -5)$, $\boldsymbol{\sigma}_0^2 = (1, 1)$. These shrinkage methods required two starting values. The initial values that we used in this simulation are $\boldsymbol{\beta}_{01(initial)} = \boldsymbol{\beta}_0 + \mathbf{2}$, $\boldsymbol{\beta}_{02(initial)} = \boldsymbol{\beta}_0 - \mathbf{2}$, $\pi_{01} = (0.5, 0.5)$, $\pi_{02} = (0.3, 0.7)$, $\boldsymbol{\sigma}_{01}^2 = (1, 2)$ and $\boldsymbol{\sigma}_{02}^2 = (3, 5)$ and all these values are fixed for all methods (ML, Ridge and Liu).

In order to investigate the performance of three methods, we used the sum of squared errors (SSE) of the paramater estimates over the 2000 replications, which is given by

$$\text{SSE}(\widehat{\boldsymbol{\Psi}}^{(m)}) = \left[(\widehat{\boldsymbol{\Psi}}^{(m)} - \boldsymbol{\Psi}_0)^\top (\widehat{\boldsymbol{\Psi}}^{(m)} - \boldsymbol{\Psi}_0)\right] \tag{5.1}$$

where $\widehat{\boldsymbol{\Psi}}^{(m)} = \{\widehat{\boldsymbol{\beta}}^{(m)}, \widehat{\pi}^{(m)}, \widehat{\sigma^2}^{(m)}\}$ for $m = 1, \cdots, 2000$. To compute the prediction performance, we use the root mean square error of prediction as follows

$$\text{MRSEP} = \frac{1}{2000} \sum_{m=1}^{2000} \text{RMSEP}^{(m)}, \tag{5.2}$$

where $\text{RMSEP}^{(m)}$ is the root mean-squared error of prediction of the m-th replication based on $K-$ fold cross validation, which is given by

$$\text{RMSEP}^{(m)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \widehat{y}_i^{(m)}\right)^2}, \tag{5.3}$$

where $\widehat{y}_i$ is the predicted response of $i-$th observation in the m-th replication.

We computed the estimation and prediction measures for the ML, Ridge and LT methods as follows. Because we shall use the cross-validation in the analysis of the mixture of linear regression, we generated sample sizes $n = \{60, 100\}$ (larger

than the sample sizes used in the first simulation study of Subsection 5.1.1) from the underlying mixture of regression models as described in equation (3.1). We then used the EM, CEM and SEM algorithms to estimate the parameters of the mixture population via ML, Ridge and LT methods. We applied the idea of K = 5 cross-validation to assess the prediction performance of the methods. To this end, we divide the data into K folds of equal sizes. We used $K-1$ folds for training and the remaining fold for prediction. We repeated the procedure for all $k = 1, \cdots, K$ to compute the corresponding value for RMSEP$^{(m)}$. Eventually, we replicated the entire procedure $m = 2000$ times and computed the median and 95% intervals of the SSE and RMSEP measures. The lower and upper bounds of the intervals correspond to 2.5 and 97.5 percentiles of 2000 replications, respectively.

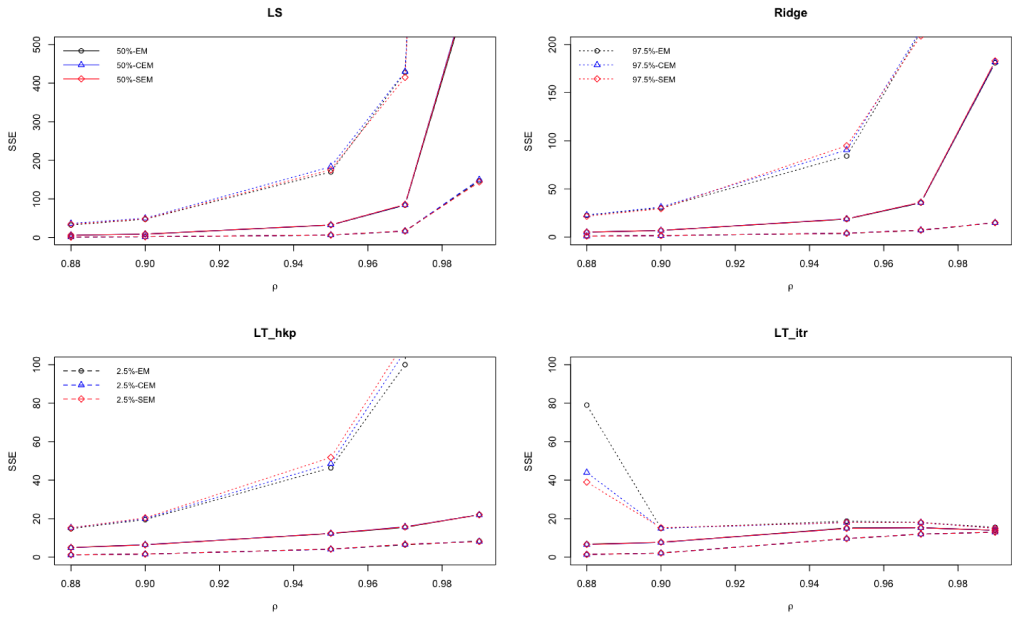

Figure 5.1: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\beta}}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of coefficients of the mixture of two regressions when $n = 60$.

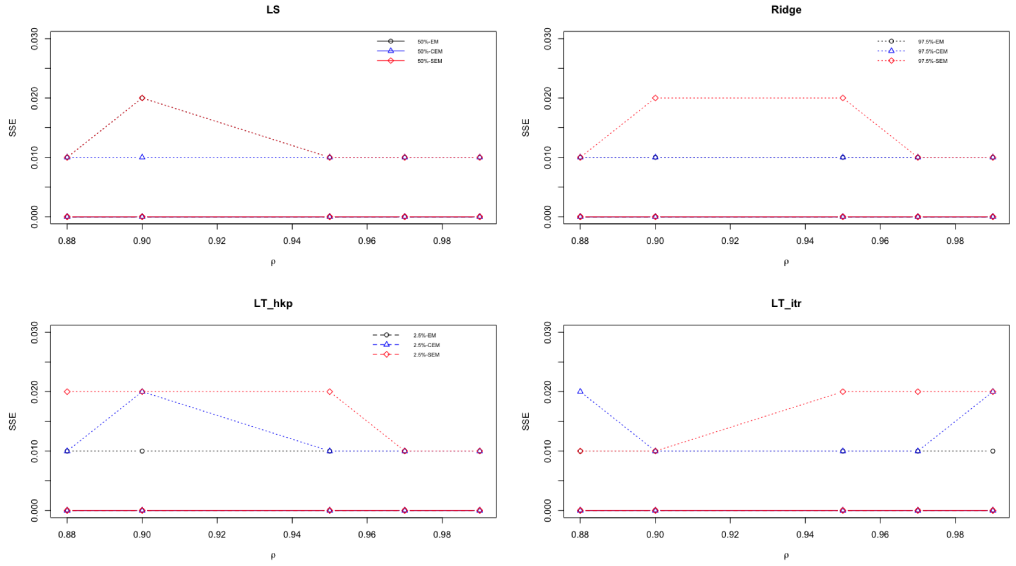

Figure 5.2: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $(\widehat{\pi})$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of the mixing proportions of the mixture of two regressions when $n = 60$.

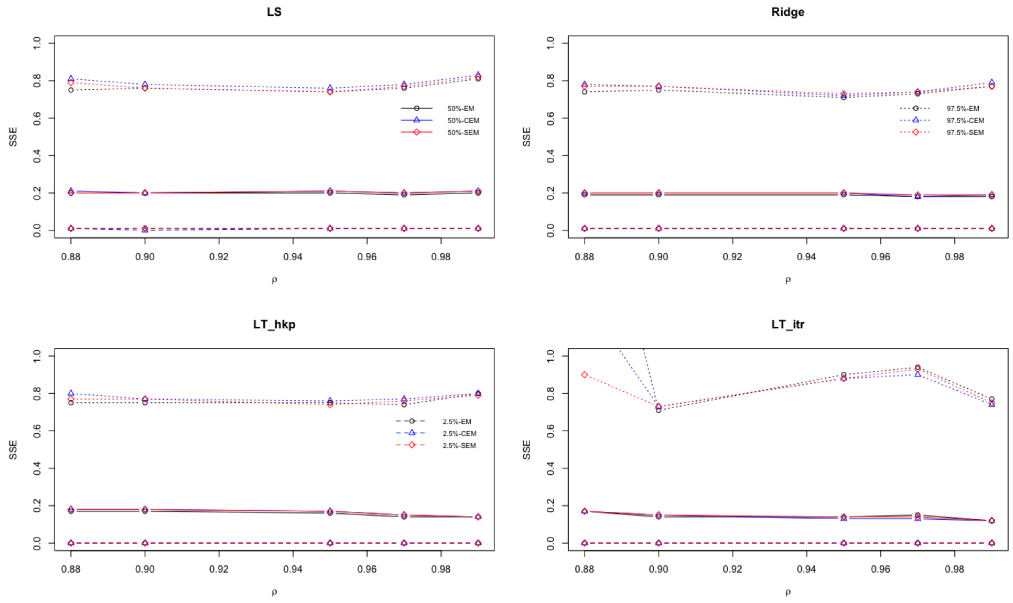

Figure 5.3: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\sigma}^2}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of the variance of error term of the mixture of two regressions when $n = 60$.

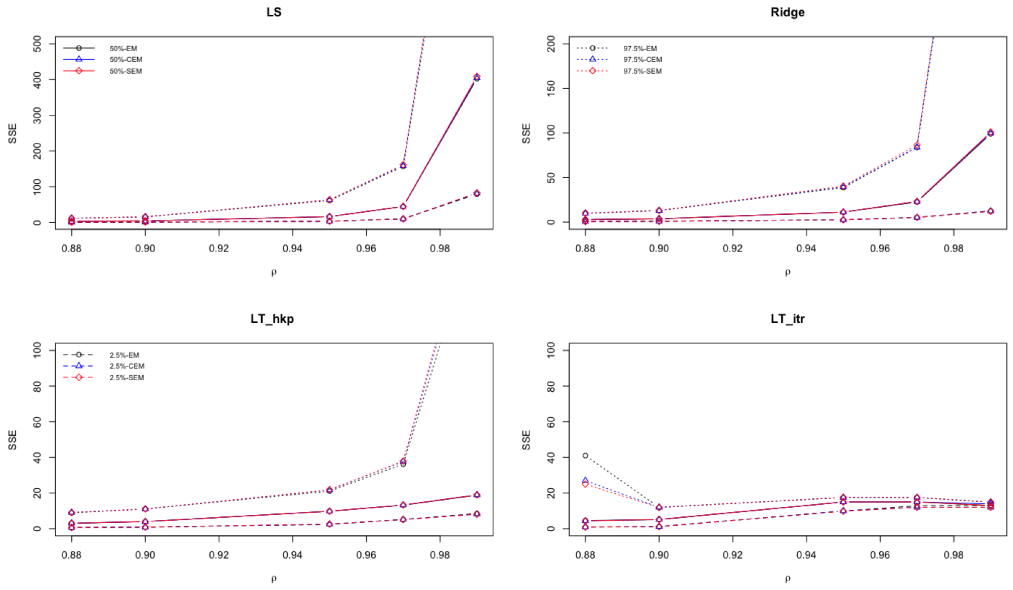

Figure 5.4: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\beta}}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of coefficients of the mixture of two regressions when $n = 100$.

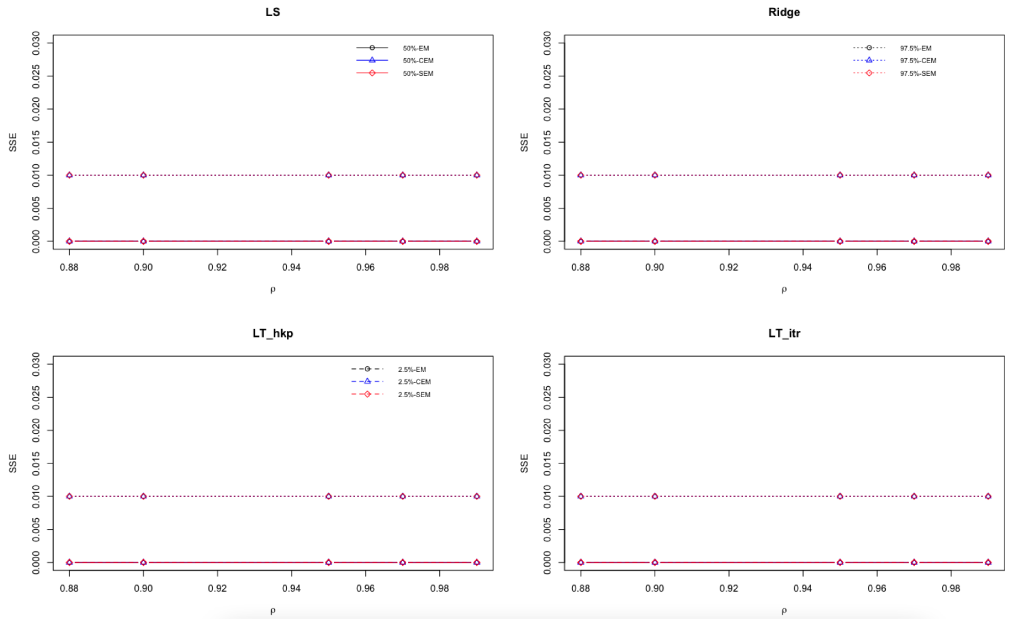

Figure 5.5: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $(\widehat{\pi})$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of the mixing proportions of the mixture of two regressions when $n = 100$.

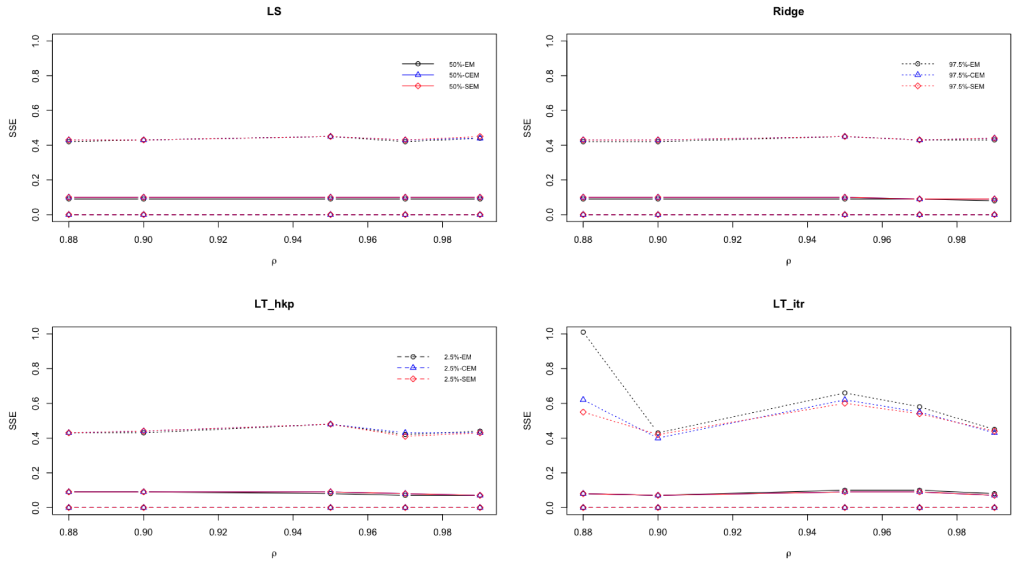

Figure 5.6: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\sigma}^2}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of the variance of error term of the mixture of two regressions when $n = 100$.

Table 5.12: The median (M) and the length of 95% intervals for the RMSEP of the ML, Ridge and LT methods in prediction of the mixture of two regressions when $n = 60$

		$\rho = 0.88$		$\rho = 0.90$		$\rho = 0.95$		$\rho = 0.97$		$\rho = 0.99$	
Estimator	Method	M	L	M	L	M	L	M	L	M	L
Ls	EM	16.3	10.8	16.7	11.1	17.7	12.5	18.1	12.1	18.3	12.3
	CEM	16.5	10.7	16.7	11.1	17.7	12.0	18.0	12.0	18.2	12.3
	SEM	16.3	11.3	16.7	11.4	17.6	11.9	17.9	11.9	18.2	12.4
Ridge	EM	16.5	11.0	16.7	11.2	17.9	12.1	18.1	12.0	18.3	12.0
	CEM	16.4	10.8	16.7	11.1	17.8	11.6	17.9	11.9	18.3	12.7
	SEM	16.4	11.0	16.7	11.5	17.7	11.6	18.0	11.8	18.3	12.3
LT(HKP)	EM	16.4	10.6	16.6	11.1	17.7	11.7	18.1	12.5	18.3	12.0
	CEM	16.4	11.3	16.6	11.2	17.6	11.9	18.1	11.6	18.3	12.2
	SEM	16.4	10.9	16.6	11.2	17.8	12.0	18.0	12.1	18.3	12.1
LT(ITE)	EM	16.4	10.8	16.8	10.9	17.3	11.4	17.7	11.8	18.1	12.0
	CEM	16.5	10.9	16.7	10.9	17.2	11.7	17.7	11.8	18.1	11.9
	ESM	16.4	11.1	16.7	11.0	17.2	11.7	17.7	12.3	18.1	11.9

Table 5.13: The median (M) and the length of 95% intervals for the RMSEP of the ML, Ridge and LT methods in prediction of the mixture of two regressions when $n = 100$

		$\rho = 0.88$		$\rho = 0.90$		$\rho = 0.95$		$\rho = 0.97$		$\rho = 0.99$	
Estimator	Method	M	L	M	L	M	L	M	L	M	L
Ls	EM	16.6	8.2	16.9	8.8	17.7	9.0	18.1	8.9	18.5	9.3
	CEM	16.5	8.6	16.8	8.7	17.6	9.4	18.1	9.0	18.5	9.9
	SEM	16.6	8.5	16.9	8.9	17.7	9.0	18.0	9.4	18.5	9.8
Ridge	EM	16.6	8.8	16.8	8.6	17.7	9.1	18.1	9.2	18.6	9.4
	CEM	16.5	8.6	16.8	8.6	17.6	9.4	18.0	9.4	18.4	9.3
	SEM	16.5	8.7	16.7	8.5	17.8	9.0	18.0	9.3	18.5	9.4
LT(HKP)	EM	16.5	8.6	16.8	8.7	17.6	9.0	18.1	8.9	18.5	9.9
	CEM	16.4	8.8	16.8	8.5	17.6	8.9	18.1	9.4	18.4	9.6
	SEM	16.5	8.6	16.9	8.8	17.7	8.9	18.0	9.2	18.4	9.6
LT(ITE)	EM	16.6	8.5	16.9	8.4	17.5	9.3	17.9	9.3	18.2	9.0
	CEM	16.5	8.6	16.8	8.8	17.3	9.0	17.8	9.2	18.2	9.3
	ESM	16.6	8.4	16.8	8.5	17.4	8.8	17.8	9.7	18.2	9.5

Figures 5.1-5.6 show the results of the simulation study with sample sizes 60 and 100. In each graph, we represent the performance of the three estimators (LS, Ridge, and Liu). For each estimator, we represent the performance in three lines, including the median of SSE, 2.5% lower-bound of SSE, and 97.5% upper-bound of SSE. We represented these nine lines in three legends in each figure.

The ML methods estimate slightly better the mixing proportion than the Ridge and LT methods. This is based on the fact that the Ridge and LT estimators are biased shrinkage methods where a slight bias is incorporated into the estimation to encounter the multicollinearity problem. We observe that the multicollinearity significantly affects the ML estimates of the coefficients and results in extremely unreliable estimates for all EM, CEM, and SEM algorithms. In contrast to ML estimates, the performance of the shrinkage approaches in estimating the coefficients of the component regressions shows a significant improvement. In the multicollinearity, the LT approaches seem to be more reliable than their Ridge equivalents. From a comparison between LT(ITR) and LT(HKP), we see that LT(HKP) provides more reliable estimates for σ^2. Among the LT(HKP) estimators, the CEM algorithm almost always outperforms its EM ad SEM counterparts. Tables 5.12 and 5.13 show the median and 95% intervals of the RMSEP for all the developed methods. The tables clearly show that all methods and EM algorithms have nearly identical prediction performances. This finding is consistent with Inan and Erdogan (2013) and Ghanem et al. (2022a) that multicollinearity seriously affects the estimation of the methods while prediction levels stay almost the same.

5.2.2 Simulation Study 2

The second simulation examines how well the estimation methods perform when the population consists of three regression models with two covariates. Assuming the correlation level $\phi = \{0.9, 0.92, 0.95, 0.97, 0.99\}$, we generated the covariates and responses as described above 5.2.1 from the mixture population when $\pi_0 = (0.3, 0.4, 0.3)$, $\boldsymbol{\sigma}_0^2 = (0.25, 1, 0.09)$ and $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{02}, \boldsymbol{\beta}_{03})$ with $\boldsymbol{\beta}_{01} = (1, 3, 4)$, $\boldsymbol{\beta}_{02} = (-1, -1, -2)$ and $\boldsymbol{\beta}_{03} = (-3, 1, -4)$. These shrinkage methods required two starting values. The initial values that we used in this simulation are $\boldsymbol{\beta}_{01(initial)} = \boldsymbol{\beta}_0 - \mathbf{1}$, $\boldsymbol{\beta}_{02(initial)} = \boldsymbol{\beta}_0 + \mathbf{1}$, $\pi_{01} = (0.3, 0.4, 0.3)$, $\pi_{02} = (0.3, 0.4, 0.4)$, $\boldsymbol{\sigma}_{01}^2 = \boldsymbol{\sigma}_0^2 + 0.02$ and $\boldsymbol{\sigma}_{02}^2 = \boldsymbol{\sigma}_0^2 + 0.02$. All of these values a little bit close from true parameters and all are fixed for all methods (ML, Ridge and Liu).

Similar to the settings of the first simulation study, we replicated 2000 times all the estimation and prediction procedures under the EM, CEM, and SEM algorithms and computed the median and 95% interval for the SSEs and RMSEP for size sizes $n = \{60, 100\}$. Figures 5.7 - 5.12 and Tables 5.14 - 5.15 report the median (M), 95% intervals for SSE and RMSEP. Here, we also observe that the ML method slightly better estimates the mixing proportions; however, the ML method results in extremely unreliable estimates for the coefficients of component regressions. It is easy to see that shrinkage estimators do better in estimating the component regression parameters. Moreover, the LT(HKP) almost always outperforms other methods and provides a more reliable estimate of the mixture of regression models. Therefore, the LT(HKP) method based on the CEM algorithm is recommended to fit the mixture of linear regression models in multicollinearity.

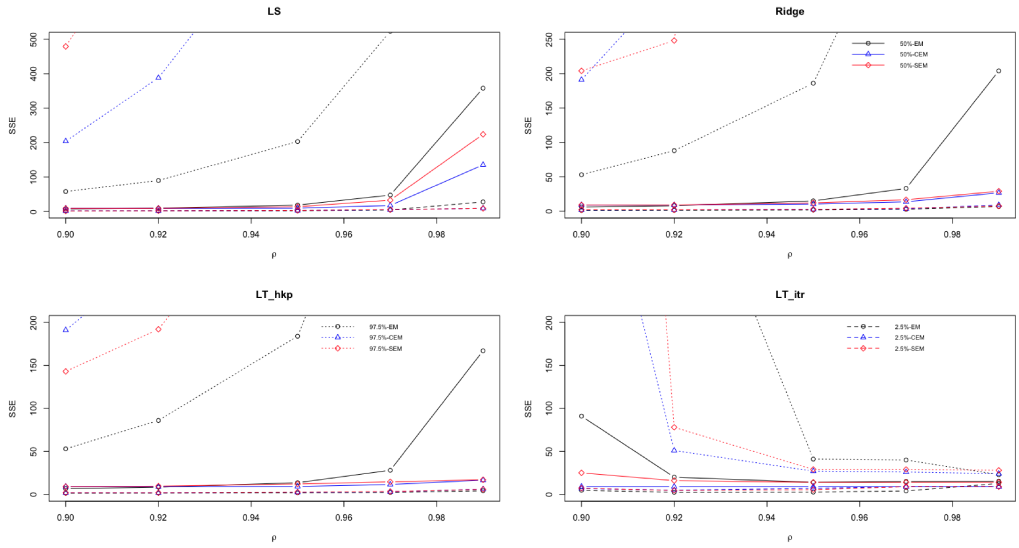

Figure 5.7: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\beta}}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of coefficients of the mixture of three regressions when $n = 60$.

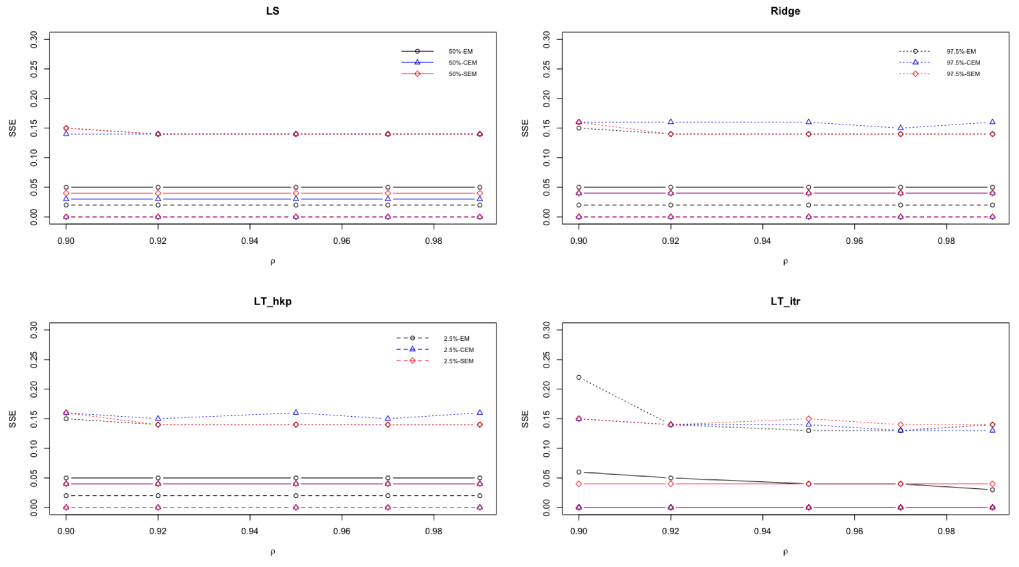

Figure 5.8: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $(\widehat{\pi})$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of mixing proportions of the mixture of three regressions when $n = 60$.

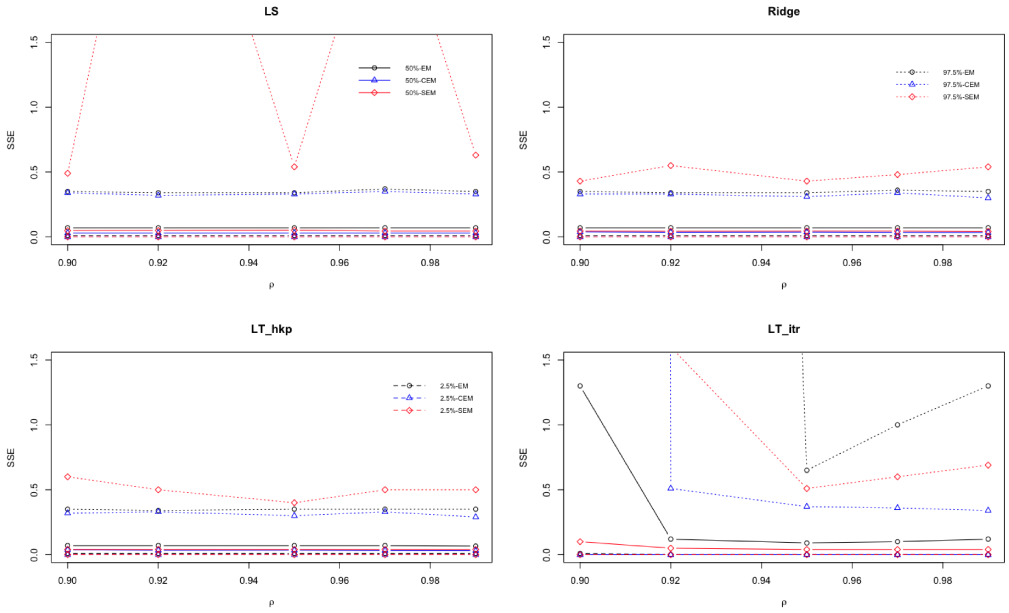

Figure 5.9: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\sigma}^2}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of the variance of error term of the mixture of three regressions when $n = 60$.

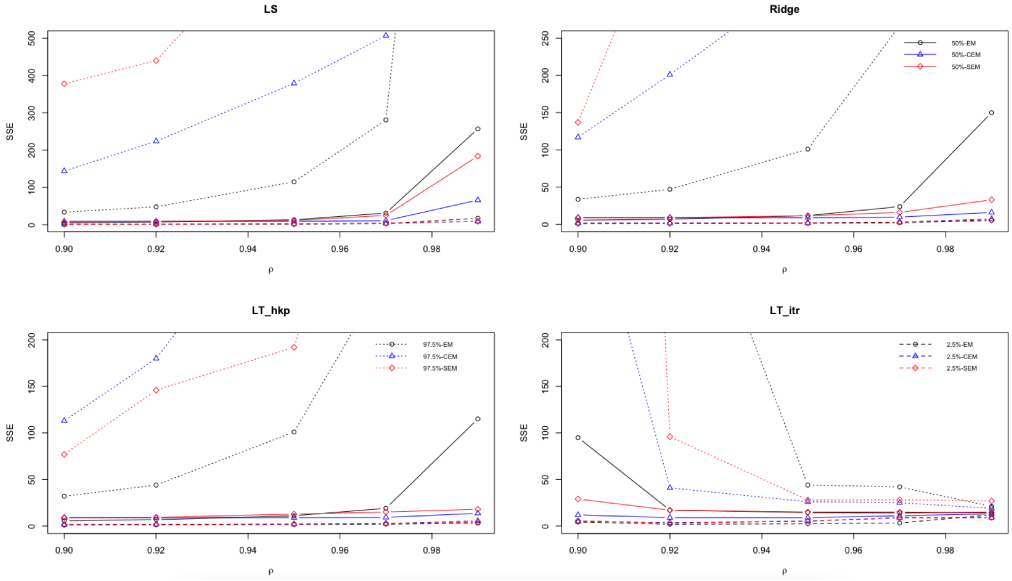

Figure 5.10: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\beta}}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of coefficients of the mixture of three regressions when $n = 100$.

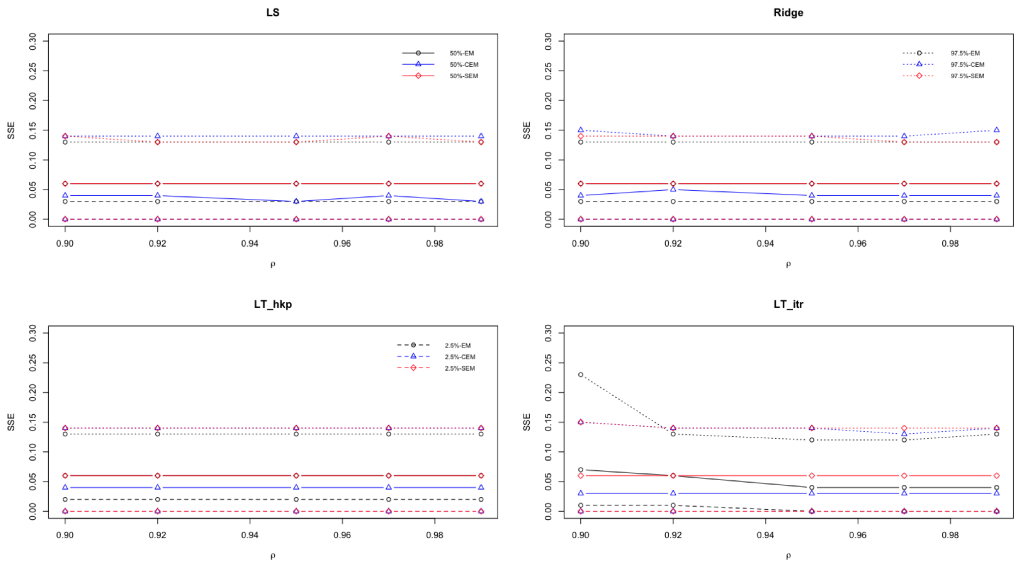

Figure 5.11: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $(\widehat{\pi})$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of mixing proportions of the mixture of three regressions when $n = 100$.

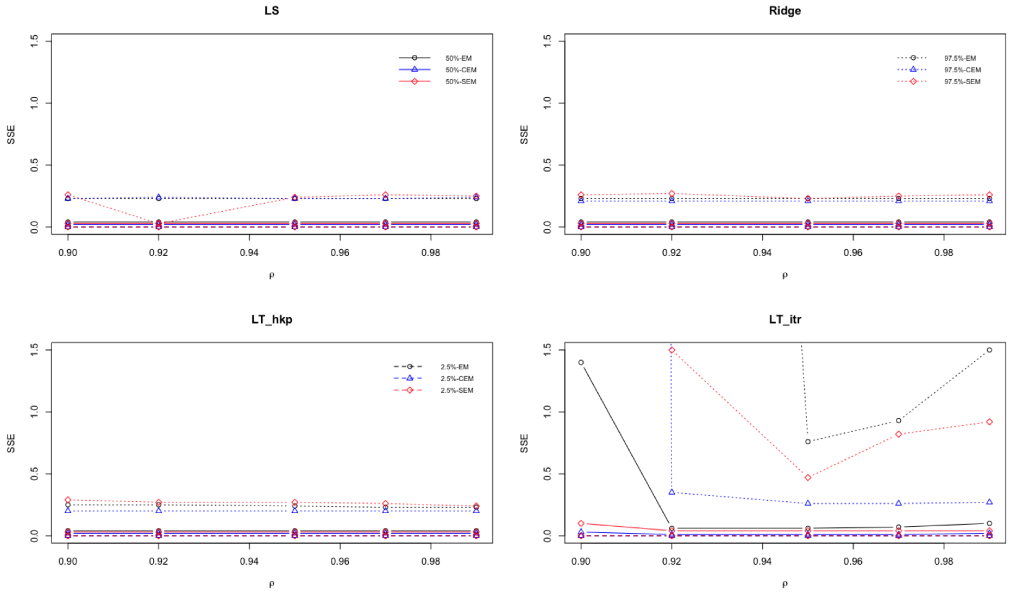

Figure 5.12: The median (solid line), lower (dotted line) and upper (dashed line) bounds of 95% intervals for SSE $\left(\widehat{\boldsymbol{\sigma}^2}\right)$ of EM (black), CEM (blue), and SEM (red) approaches in the estimation of the variance of error term of the mixture of three regressions when $n = 100$.

Table 5.14: The median (M) and the length of 95% intervals for the RMSEP of the ML, Ridge and LT methods in prediction of the mixture of three regressions when $n = 60$

		$\rho = 0.90$		$\rho = 0.92$		$\rho = 0.95$		$\rho = 0.97$		$\rho = 0.99$	
Estimator	Method	M	L	M	L	M	L	M	L	M	L
Ls	EM	6.0	4.0	6.1	3.9	6.2	4.1	6.4	4.1	6.4	4.3
	CEM	6.1	3.9	6.2	4.0	6.4	4.3	6.5	4.3	6.6	4.4
	SEM	6.1	3.9	6.2	4.1	6.3	4.2	6.4	4.2	6.5	4.3
Ridge	EM	6.0	4.0	6.1	3.9	6.2	4.1	6.3	4.2	6.4	4.5
	CEM	6.1	4.0	6.2	4.1	6.3	4.2	6.5	4.2	6.6	4.3
	SEM	6.0	4.0	6.2	4.2	6.4	4.1	6.5	4.2	6.5	4.3
LT(HKP)	EM	5.9	3.9	6.0	3.9	6.2	4.2	6.3	4.2	6.4	4.1
	CEM	6.0	4.1	6.1	4.0	6.3	4.3	6.5	4.3	6.5	4.4
	SEM	6.0	4.0	6.1	4.0	6.3	4.3	6.5	4.2	6.6	4.4
LT(ITE)	EM	6.5	6.9	6.1	3.9	6.1	4.1	6.3	4.0	6.3	4.1
	CEM	6.3	4.4	6.3	4.0	6.4	4.2	6.5	4.2	6.6	4.3
	SEM	6.2	4.5	6.2	4.0	6.3	4.1	6.4	4.2	6.5	4.3

Table 5.15: The median (M) and the length of 95% intervals for the RMSEP of the ML, Ridge and LT methods in prediction of the mixture of three regressions when $n = 100$

		$\rho = 0.90$		$\rho = 0.92$		$\rho = 0.95$		$\rho = 0.97$		$\rho = 0.99$	
Estimator	Method	M	L	M	L	M	L	M	L	M	L
Ls	EM	5.9	3.0	6.1	3.2	6.3	3.1	6.4	3.4	6.5	3.4
	CEM	6.1	3.0	6.3	3.1	6.4	3.3	6.5	3.3	6.6	3.4
	SEM	6.1	3.0	6.1	3.2	6.4	3.3	6.5	3.1	6.5	3.4
Ridge	EM	6.0	3.1	6.1	3.1	6.3	3.3	6.4	3.2	6.5	3.3
	CEM	6.1	3.0	6.2	3.0	6.4	3.2	6.5	3.4	6.6	3.4
	SEM	6.0	3.0	6.2	3.1	6.4	3.2	6.5	3.3	6.6	3.5
LT(HKP)	EM	5.9	3.1	6.0	3.0	6.2	3.2	6.4	3.2	6.5	3.3
	CEM	6.0	3.2	6.2	3.2	6.3	3.2	6.5	3.4	6.6	3.4
	SEM	6.0	3.1	6.1	3.1	6.9	3.1	6.5	3.2	6.5	3.4
LT(ITE)	EM	6.4	6.2	6.1	3.0	6.2	3.1	6.3	3.0	6.3	3.5
	CEM	6.3	3.3	6.3	3.2	6.4	3.3	6.5	3.4	6.6	3.6
	ESM	6.2	4.2	6.2	3.0	6.3	3.1	6.3	3.2	6.4	3.4

5.3 Bone Data Analysis

Osteoporosis is a bone disorder that arises when the body's bone architecture dramatically reduces. This deterioration causes a slew of serious health problems. Patients with osteoporosis, for example, are more susceptible to skeletal fragility and fractures in areas such as the vertebral, hip, and femur (Cummings et al., 1993; Melton III et al., 1998). Osteoporosis has a significant impact on the health and survival of a patient. More than half of patients with osteoporotic hip fractures are unable to live independently, and one-third of these patients will die within one year as a result of the disease's medical complications (Bliuc et al., 2009; Neuburger et al., 2015). The financial load of osteoporosis is also undeniable on community health. For example, according to (Lim et al., 2016), the annual cost of osteoporosis and its linked health problems in South Korea is twice that of diabetes.

Bone mineral density (BMD) is the most influential factor in diagnosing osteoporosis, according to a WHO expert panel (WHO, 1994). When the BMD score is fewer than 2.5 SDs from the BMD norm, it is diagnosed as osteoporosis (i.e., the mean of BMD scores of healthy individuals between 20-29). The density of bone tissues grows until the age range of 20-30, after which it declines as the person ages. Aside from age, other research studies in the literature looked at the relationship between osteoporosis and patient characteristics like sex, weight, and BMI (Felson et al., 1993; Kim et al., 2012).

5.3.1 Bone Data Analysis For Mixture of Logistic Regression

Even though measuring BMD scores is costly, practitioners have access to a variety of easily accessible patient information, such as physical and demographic characteristics and BMD results from previous surveys. Logistic regression is a useful statistical approach for using these factors to explain patient statistics on osteoporosis. Within each osteoporosis class, the impact of these factors can differ. As a result, in an unsupervised learning technique, a mixture of logistic regressions can be used to estimate the effects of various characteristics. The data on bone mineral density from the National Health and Nutritional Examination Survey was used in this numerical investigation (NHANES III). Between 1988 and 1994, the Centers for Disease Control and Prevention (CDC) conducted a study of 33999 American adults. A total of 182 women aged 50 and up took part in two bone examinations, where there are 36 bone characteristics available for each individual. We treated these 182 women as our underlying population because of the severe impact of osteoporosis on the elderly female population. We took the total BMD from the second bone examination and converted it to a binary osteoporosis status as our response variable. To covert our response variable (BMD) to a binary status, we first calculated the mean $\left(\overline{BMD}_R\right)$ and standard deviation sd(BMD_R) for the BMD values for the reference group (i.e, women aged between 20 to 30) and computed the BMD norm $m_0 = \left(\overline{BMD}_R\right) - sd\,(BMD_R)$. We then compared the BMD value of individuals with the BMD norm; if it is greater than m_0, the BMD status of the individual is assigned as 1; otherwise 0. We also used two easy-to-measure physical features as covariates in the logistic regressions: arm and bottom circumferences.

The high association between the covariates $\rho = 0.81$ indicates the multicollinearity problem in the mixture of logistic regressions. We applied R package mixtools to all the 182 observations and found the estimates of the coefficients and mixing proportions. We then treated these estimates as a true parameter of the bone population in this real data study. We replicated 2000 times the ML, Ridge, and LT methods in estimating the parameters of the bone mineral population with training sample size $n = \{20, 40, 80, 100\}$ and test sample size (taken independently from the training step) of size 50. We then computed the estimation and predication measures $\sqrt{\text{SSE}(\widehat{\boldsymbol{\beta}})}$, $\sqrt{\text{SSE}(\widehat{\pi})}$, Error, Sensitivity, Specificity as described in Section 5.1.1 where $\boldsymbol{\beta}_0$ and π_0 are obtained by ML estimates of the parameters using the complete information of the population.

Table 5.16 shows the median (M) and 95% intervals of the above estimation and prediction measures. The lower (L) and upper (U) bounds of the intervals were determined by 2.5 and 97.5 percentiles of the estimates. While the ML method slightly estimates the mixing proportions better, the ML method becomes extremely unreliable in estimating the coefficients of component logistic regressions. Unlike the ML, the Ridge and shrinkage methods could handle the multicollinearity issue in the estimation problem. Comparing the shrinkages methods, $\widehat{\boldsymbol{\beta}}_{LT}$ significantly outperforms $\widehat{\boldsymbol{\beta}}_{R}$ in estimating the coefficients of the mixture. Therefore, the LT shrinkage method is recommended to estimate a mixture of logistic regressions when there is multicollinearity in bone mineral data.

Table 5.16: The median (M), lower (L) and upper (U) bounds of 95% intervals for $\sqrt{\text{SSE}}$, Error, Sensitivity (SN) and Specificity (SP) of the methods in the analysis of bone mineral data with sample size $n = \{20, 40, 80, 100\}$.

			$\sqrt{\text{SSE}}$			Error			SN			SP		
n	SEM	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U	M	L	U
20	ML	β	7.8	.70	6×10^{51}	.36	.20	.64	.00	.00	.81	1	.19	1
		π	.1	.00	.5									
	Ridge	β	1.9	.4	30.2	.44	.26	.66	.35	.00	.84	.66	.18	1
		π	.2	.00	.7									
	LT	β	1.9	.58	3.2	.46	.28	.62	.33	.00	.78	.64	.28	.97
		π	.3	.00	.7									
40	ML	β	8.2	1.3	4×10^{64}	.34	.20	.56	.00	.00	.61	1	.41	1
		π	.07	.00	.45									
	Ridge	β	1.8	1.2	26.4	.44	.28	.62	.33	.00	.75	.67	.33	.97
		π	.22	.025	.7									
	LT	β	1.9	0.6	2.1	.46	.30	.60	.35	.07	.69	.64	.38	.89
		π	.3	.05	.7									
80	ML	β	9.6	2.0	5×10^{70}	.34	.20	.58	.00	.00	.63	1	.42	1
		π	.05	.00	.42									
	Ridge	β	1.8	1.2	8.9	.44	.30	.60	.33	.00	.67	.68	.45	1
		π	.25	.025	.67									
	LT	β	1.9	1.05	2.0	.46	.30	.60	.33	.07	.63	.66	.44	.87
		π	.27	.025	.7									
100	ML	β	8.9	2.0	2×10^{73}	.34	.20	.56	.00	.00	.64	1	.41	1
		π	.04	.00	.4									
	Ridge	β	1.8	.96	8.5	.44	.28	.60	.32	.00	.64	.67	.44	.97
		π	.26	.02	.67									
	LT	β	1.9	1.1	2.0	.46	.30	.60	.33	.08	.64	.65	.45	.85
		π	.27	.03	.7									

5.3.2 Bone Data Analysis For Mixture of Linear Regression Models

In the regressions, we additionally utilized two easy-to-measure physical traits as covariates: circumferences of the arms and the bottom. The high association between the covariates $\rho = 0.81$ indicates the multicollinearity problem in the mixture of regressions. We replicated 2000 Monte Carlo simulations for the ML, Ridge, and LT methods in estimating the parameters of the bone mineral population with sample size $n = \{60, 100\}$. We used 5-fold cross-validation where, in each iteration, we used 4-fold for training and one remaining fold for testing. Then we changed this testing fold iteratively to cover all sample sizes, to investigate the estimation and prediction performance for all methods. We then computed the estimation and prediction measures $\sqrt{\text{SSE}(\widehat{\boldsymbol{\beta}})}$, $\sqrt{\text{SSE}(\widehat{\pi})}$, $\sqrt{\text{SSE}(\widehat{\sigma^2})}$ and MRSEP as described in Section 5.2 where $\boldsymbol{\beta}_0$, π_0 and σ_0^2 are obtained by ML estimates of the parameters using the complete information of the population. Similar to Subsection 5.3.1, we applied R package mixtools to all the 182 observations and found the estimates of the coefficients and mixing proportions. We then used these estimates as a true parameter of the bone population in this real data study

Table 5.17: The median (M), lower (L) and upper (U) bounds of 95% intervals for $\sqrt{\text{SSE}}$ of the methods in the analysis of bone mineral data with sample size $n = 60$.

		CEM			SEM			EM		
Methods	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U
LS	β	.010	.002	.165	.019	.003	.213	.018	.003	.134
	π	.333	.100	.366	.333	.183	.366	.218	.015	.365
	σ^2	.003	.000	.014	.006	..000	.014	.004	.000	.014
Ridge	β	.009	.002	.165	.013	.002	.166	.012	.002	.118
	π	.333	.100	.366	.333	.166	.366	.214	.019	.366
	σ^2	.003	.000	.014	.006	.000	.014	.004	.000	.014
LT(HKP)	β	.009	.002	.165	.010	.003	.183	.010	.003	.067
	π	.333	.100	.366	.333	.150	.366	.205	.013	.372
	σ^2	.003	.000	.014	.006	.000	.014	.004	.000	.016
LT(ITE)	β	.009	.002	.010	.009	.006	.011	.009	.007	.010
	π	.300	.100	.366	.350	.116	.566	.575	.032	.599
	σ^2	.002	.000	.014	.005	.000	.014	.003	.000	.009

Table 5.18: The median (M), lower (L) and upper (U) bounds of 95% intervals for $\sqrt{\text{SSE}}$ of the methods in the analysis of bone mineral data with sample size $n = 100$.

		CEM			SEM			EM		
Methods	$\boldsymbol{\Psi}$	M	L	U	M	L	U	M	L	U
LS	β	.010	.002	.126	.014	.003	.202	.014	.002	.112
	π	.350	.100	.380	.360	.210	.380	.222	.016	.370
	σ^2	.005	.000	.014	.006	..000	.014	.003	.000	.014
Ridge	β	.009	.002	.123	.011	.003	.165	.009	.002	.086
	π	.350	.100	.380	.360	.210	.380	.220	.019	.370
	σ^2	.005	.000	.014	.006	.000	.014	.003	.000	.014
LT(HKP)	β	.009	.002	.123	.010	.003	.133	.010	.002	.047
	π	.350	.100	.380	.360	.190	.380	.207	.019	.370
	σ^2	.004	.000	.014	.005	.000	.014	.003	.000	.014
LT(ITE)	β	.009	.002	.010	.009	.007	.010	.009	.007	.009
	π	.310	.100	.380	.360	.150	.580	.584	.040	.600
	σ^2	.004	.000	.014	.005	.000	.014	.002	.000	.007

Table 5.19: The median (M), the lower and the upper of 95% intervals for the RMSEP of the ML, ridge and LT methods in prediction of the Bone real data

		$n = 60$			$n = 100$		
Estimator	Method	M	L	U	M	L	U
LS	EM	.139	.105	.195	.135	.110	.173
	CEM	.149	.114	.220	.141	.113	.190
	SEM	.140	.104	.234	.137	.109	.205
Ridge	EM	.137	.104	.186	.133	.109	.172
	CEM	.148	.115	.207	.139	.113	.189
	SEM	.140	.103	.223	.136	.109	.194
LT(HKP)	EM	.135	.104	.182	.132	.109	.168
	CEM	.148	.115	.208	.139	.113	.188
	SEM	.139	.104	.221	.137	.109	.194
LT(ITE)	EM	.125	.101	.155	.124	.104	.145
	CEM	.153	.119	.193	.146	.117	.181
	SEM	.143	.110	.186	.140	.113	.171

Tables 5.17-5.19 show the median, the lower and the upper of 95% intervals for the $\sqrt{\text{SSE}}$ and RMSEP of the ML, ridge and LT methods in the estimation and prediction of the Bone real data. The lower (L) and upper (U) bounds of the CIs were determined by 2.5 and 97.5 percentiles of the estimates. Although all ML, Ridge and LT methods almost perform identically in estimating the mixing proportion and component variances, $\widehat{\boldsymbol{\beta}}_{ML}$ become considerably unreliable. Unlike ML methods, the LT and Ridge shrinkage methods could appropriately handle the multicollinearity in estimating the coefficients of component regressions. Comparing the shrinkage methods, we observe that the LT estimators appear more reliable than their Ridge counterparts in estimating the parameters of the bone mineral population.

Chapter 6

Summary and Concluding Remarks

6.1 Summary

Many medical applications, such as osteoporosis research, require a costly and time-consuming method to diagnose the disease status; however, practitioners have access to various easy-to-measure patient variables, such as physical and demographic information. Logistic regression is a robust statistical tool for utilizing these features to explain an illness's status. In an unsupervised learning technique, a mixture of logistic regressions can be used to study the effect of covariates on the binary response, and a mixture of regressions can be utilized in an unsupervised learning technique to explore the effect of covariates on the response when the population contains different subpopulations.

This thesis investigated the estimation of the parameters on the mixture of logistic and mixture of linear regression models in the presence of multicollinearity. We developed Liu-type (LT) shrinkage estimator for finite mixture models (FMMs)

to deal with the multicollinearity problem.

Although the maximum likelihood (ML) method is the usual method for estimating the parameter of a mixture of logistic regressions and mixture of linear regression models, multicollinearity significantly impacts ML estimates. The properties of the Ridge and LT shrinkage methods in estimating the mixture of logistic regressions and the mixture of regression models were examined in this study. Our proposed methods are biased, and we only recommend them when there is multicollinearity.

According to mixture of linear regressions and mixture of logistic regression models, we discovered that the ML technique estimates the mixing proportions of the mixture models slightly better than shrinkage methods based on extensive numerical simulations. Because shrinkage estimators are biased methods, they are meant to overcome the ill-conditioned design matrix at the cost of bias in the estimation. With multicollinearity, the ML approach for predicting mixture coefficients becomes significantly unreliable. Unlike the ML method, the proposed shrinkage approaches give more reliable estimations. When comparing shrinkage techniques, $\widehat{\boldsymbol{\beta}}_{LT}$ outperforms considerably $\widehat{\boldsymbol{\beta}}_{R}$ in the presence of multicollinearity in the mixture of regression models and a mixture of logistic regressions.

Although the proposed shrinkage method performed well in dealing with multicollinearity problems, the performance of the methods was only assessed based on the SEM algorithm in the mixture of Logistic regression. Note that the convergence rate of the shrinkage methods under CEM and EM approaches reduces significantly. For this reason, we only reported the results based on the SEM approach in this thesis. We believe that this computational issue arises because one of the logistic regression components becomes empty in the early iterations of the CEM and EM

algorithm.

Finally, we applied the proposed methods to bone mineral data to analyze the bone disorder status of women aged 50 and older. When the correlation between covariates is small, and hence the collinearity is not severe, these shrinkage methods are not recommended as they result in biased estimates.

6.2 Future Work

In the future, we plan to work on count data, a statistical data type that describes countable quantities using only count values. Poisson distribution is one of the most important distributions that deal with counting data. We will focus on mixture of Poisson regression model, so we shall study the shrinkage estimators to deal with multicollinearity problem in the problem of mixture of Poisson regression models.

Bibliography

Mahdi A Alkhamisi and Ghazi Shukur. Developing ridge parameters for sur model. *Communications in Statistics—Theory and Methods*, 37(4):544–564, 2008.

Paul D Allison. Comparing logit and probit coefficients across groups. *Sociological Methods & Research*, 28(2):186–208, 1999.

Anestis Antoniadis. Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, 6(2):97–130, 1997.

Ronald G Askin. Multicollinearity in regression: Review and examples. *Journal of Forecasting*, 1(3):281–292, 1982.

Xiuqin Bai, Weixin Yao, and John E Boyer. Robust fitting of mixture regression models. *Computational Statistics & Data Analysis*, 56(7):2347–2359, 2012.

Sivaraman Balakrishnan, Martin J Wainwright, and Bin Yu. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120, 2017.

Francesco Bartolucci and Luisa Scaccia. The use of mixtures for dealing with non-normal regression errors. *Computational Statistics & Data Analysis*, 48(4):821–834, 2005.

M Beer. Asymptotic properties of the maximum likelihood estimator in dichotomous logistic regression models. *Switzerland: University of Fribourg Switzerland*, 2001.

David A Belsley, Edwin Kuh, and RE Welsch. Identifying influential data and sources of collinearity. *Regression Diagnostics*, 1980.

Dana Bliuc, Nguyen D Nguyen, Vivienne E Milch, Tuan V Nguyen, and John A Eisman. Mortality risk associated with low-trauma osteoporotic fracture and subsequent fracture in men and women. *Jama*, 301(5):513–521, 2009.

Allan Bluman. *Elementary Statistics: A step by step approach 9e*. McGraw Hill, 2014.

Dankmar Böhning. *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others*, volume 81. CRC press, 1999.

Carl R Boyd, Mary Ann Tolson, and Wayne S Copes. Evaluating trauma care: the triss method. trauma score and the injury severity score. *The Journal of Trauma*, 27(4):370–378, 1987.

Richard Morrison Cassie. Some uses of probability paper in the analysis of size frequency distributions. *Marine and Freshwater Research*, 5(3):513–522, 1954.

Gilles Celeux. The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quarterly*, 2:73–82, 1985.

Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3): 315–332, 1992.

Gilles Celeux and Gerard Govaert. Comparison of the mixture and the classification maximum likelihood in cluster analysis. *Journal of Statistical Computation and Simulation*, 47(3-4):127–146, 1993.

Gilles Celeux, Didier Chauveau, and Jean Diebolt. Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, 55(4):287–314, 1996.

Hanfeng Chen and Jiahua Chen. Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, 13(2):351–366, 2003.

Jiahua Chen, Pengfei Li, and Yuejiao Fu. Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107(499):1096–1105, 2012.

Yudong Chen, Xinyang Yi, and Constantine Caramanis. A convex formulation for mixed regression with two components: Minimax optimal rates. In *Conference on Learning Theory*, pages 560–604. PMLR, 2014.

Stephen Clark. Traffic prediction using multivariate nonparametric regression. *Journal of Transportation Engineering*, 129(2):161–168, 2003.

A Clifford Cohen. Estimation in mixtures of two normal distributions. *Technometrics*, 9(1):15–28, 1967.

DavidR Cox and EJ Snell. The analysis of binary data. london: Chapman and hall. 1989.

Jan Salomon Cramer. The origins of logistic regression. 2002.

Steven R Cummings, W Browner, DM Black, MC Nevitt, HK Genant, J Cauley, K Ensrud, J Scott, and TM Vogt. Bone density at various sites for prediction of hip fractures. *The Lancet*, 341(8837):72–75, 1993.

Steven R Cummings, Michael C Nevitt, Warren S Browner, Katie Stone, Kathleen M Fox, Kristine E Ensrud, Jane Cauley, Dennis Black, and Thomas M Vogt. Risk factors for hip fracture in white women. *New England journal of medicine*, 332(12):767–774, 1995.

C De Laet, JA Kanis, Anders Odén, H Johanson, Olof Johnell, P Delmas, JA Eisman, H Kroger, S Fujiwara, P Garnero, et al. Body mass index as a predictor of fracture risk: a meta-analysis. *Osteoporosis International*, 16(11):1330–1338, 2005.

Richard D De Veaux. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, 8(3):227–245, 1989.

Lawrence T DeCarlo. Recognizing uncertainty in the q-matrix via a bayesian extension of the dina model. *Applied Psychological Measurement*, 36(6):447–468, 2012.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

José G Dias and Michel Wedel. An empirical comparison of em, sem and mcmc performance for problematic gaussian mixture likelihoods. *Statistics and Computing*, 14(4):323–332, 2004.

David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

Norman R Draper and Harry Smith. *Applied regression analysis*, volume 326. John Wiley & Sons, 1998.

Diane E Duffy and Thomas J Santner. On the small sample properties of norm-restricted maximum likelihood estimators for logistic regression models. *Communications in Statistics-Theory and Methods*, 18(3):959–980, 1989.

Gilles Dutilh, Eric-Jan Wagenmakers, Ingmar Visser, and Han LJ van der Maas. A phase transition model for the speed-accuracy trade-off in response time experiments. *Cognitive Science*, 35(2):211–250, 2011.

Ali El Zaart, Djemel Ziou, Shengrui Wang, and Qingshan Jiang. Segmentation of sar images. *Pattern Recognition*, 35(3):713–724, 2002.

Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Taiwo Stephen Fayose and Kayode Ayinde. Different forms biasing parameter for generalized ridge regression estimator. *International Journal of Computer Applications*, 181(37):2–29, 2019.

David T Felson, Yuqing Zhang, Marian T Hannan, and Jennifer J Anderson. Effects of weight and body mass index on bone mineral density in men and women: the framingham study. *Journal of Bone and Mineral Research*, 8(5):567–573, 1993.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1): 27–38, 1993.

David A Freedman. *Statistical Models: theory and practice*. Cambridge university press, 2009.

W David Furman and Bruce G Lindsay. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Computational statistics and Data analysis*, 17(5):493–507, 1994.

Scott Gaffney and Padhraic Smyth. Trajectory clustering with mixtures of regression models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 63–72, 1999.

Selvanayagam Ganesalingam. Classification and mixture approaches to clustering via maximum likelihood. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 38(3):455–466, 1989.

Elsayed Ghanem, Armin Hatefi, and Hamid Usefi. Liu-type shrinkage estimators for mixture of logistic regressions: An osteoporosis study. *arXiv preprint arXiv:2209.01731*, 2022a.

Elsayed Ghanem, Armin Hatefi, and Hamid Usefi. Unsupervised liu-type shrinkage estimators for mixture of regression models. *arXiv preprint arXiv:2209.04739*, 2022b.

Stanton A Glantz, Bryan K Slinker, and Torsten B Neilands. Primer of applied regression and analysis of variance. mcgraw-hill. *Inc., New York*, 1990.

Robert J Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951, 1992.

Bettina Grün and Friedrich Leisch. Applications of finite mixtures of regression models. *URL: http://cran. r-project. org/web/packages/flexmix/vignettes/regression-examples. pdf*, 2007.

JP Harding. The use of probability paper for the graphical analysis of polymodal frequency distributions. *Journal of the Marine Biological Association of the United Kingdom*, 28(1):141–153, 1949.

Frank E Harrell Jr, Peter A Margolis, Sandy Gove, Karen E Mason, E Kim Mulholland, Deborah Lehmann, Lulu Muhe, Salvacion Gatchalian, and Heinz F Eichenwald. Development of a clinical prediction model for an ordinal outcome: the world health organization multicentre study of clinical signs and etiological agents of pneumonia, sepsis and meningitis in young infants. *Statistics in Medicine*, 17 (8):909–944, 1998.

Trevor Hastie, Jerome Friedman, and Robert Tibshirani. Model inference and averaging. In *The Elements of Statistical Learning*, pages 225–256. Springer, 2001.

Armin Hatefi, Mohammad Jafari Jozani, and Omer Ozturk. Mixture model analysis of partially rank-ordered set samples: Age groups of fish from length-frequency data. *Scandinavian Journal of Statistics*, 42(3):848–871, 2015.

Armin Hatefi, Nancy Reid, Mohammad Jafari Jozani, and Omer Ozturk. Finite mixture modeling, classification and statistical learning with order statistics. *Statistica Sinica*, 30(4):1881–1903, 2020.

Dollena S Hawkins, David M Allen, and Arnold J Stromberg. Determining the

number of components in mixtures of linear models. *Computational Statistics & Data Analysis*, 38(1):15–48, 2001.

Georg Heinze and Michael Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419, 2002.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Arthur E Hoerl, Robert W Kannard, and Kent F Baldwin. Ridge regression: some simulations. *Communications in Statistics-Theory and Methods*, 4(2):105–123, 1975.

JP Hoffmann and K Shafer. Linear regression analysis: Applications and assumptions. 2nd, 2015.

DW Hosmer and S Lemeshow. Applied logistic regression,(john wiley & sons, inc.: New york). 1989.

Peter Howell and Stephen Davis. Predicting persistence of and recovery from stuttering by the teenage years based on information gathered at age 8 years. *Journal of Developmental & Behavioral Pediatrics*, 32(3):196–205, 2011.

Deniz Inan and Birsen E Erdogan. Liu-type logistic estimator. *Communications in Statistics-Simulation and Computation*, 42(7):1578–1586, 2013.

Alan Julian Izenman. Modern multivariate statistical techniques. *Regression, classification and Manifold Learning*, 10:978–0, 2008.

Xiaoqian Jiang, Robert El-Kareh, and Lucila Ohno-Machado. Improving predictions in imbalanced data using pairwise expanded logistic regression. In *AMIA annual symposium proceedings*, volume 2011, page 625. American Medical Informatics Association, 2011.

PN Jones and Geoffrey J McLachlan. Fitting finite mixture models in a regression context. *Australian Journal of Statistics*, 34(2):233–240, 1992.

Susan R Jones and Marylu K McEwen. A conceptual model of multiple dimensions of identity. *Journal of college student development*, 41(4):405–414, 2000.

Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994.

Ravi Kain and Ajay Verma. Logistics management in supply chain–an overview. *Materials Today: Proceedings*, 5(2):3811–3816, 2018.

Sang Jun Kim, Won-Gyu Yang, Eun Cho, and Eun-Cheol Park. Relationship between weight, body mass index and bone mineral density of lumbar spine in women. *Journal of Bone Metabolism*, 19(2):95–102, 2012.

Yongdai Kim, Sunghoon Kwon, and Seuck Heun Song. Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data. *Computational Statistics & Data Analysis*, 51(3):1643–1655, 2006.

John Kelly Kissock, Jeff S Haberl, and David E Claridge. Inverse modeling toolkit: Numerical algorithms. *ASHRAE Transactions*, 109:425, 2003.

Jason M Klusowski, Dana Yang, and WD Brinda. Estimating the coefficients of a mixture of two linear regressions by expectation maximization. *arXiv preprint arXiv:1704.08231*, 2017.

M Kutner, C Nachtsheim, J Neter, and W Li. Applied linear statistical models: Mcgraw-hill, 2004.

SQ Lafi and JB Kaneene. An explanation of the use of principal-components analysis to detect and correct for multicollinearity. *Preventive Veterinary Medicine*, 13(4): 261–275, 1992.

Deborah L Levy, Philip S Holzman, Steven Matthysse, and Nancy R Mendell. Eye tracking dysfunction and schizophrenia: a critical perspective. *Schizophrenia Bulletin*, 19(3):461–536, 1993.

Hee-Sook Lim, Soon-Kyung Kim, Hae-Hyeog Lee, Dong Won Byun, Yoon-Hyung Park, and Tae-Hee Kim. Comparison in adherence to osteoporosis guidelines according to bone health status in korean adult. *Journal of Bone Metabolism*, 23 (3):143–148, 2016.

Bruce G Lindsay. Mixture models: theory, geometry, and applications. Ims, 1995.

Kejian Liu. Using liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, 32(5):1009–1020, 2003.

Mirko Manchia, Clement C Zai, Alessio Squassina, John B Vincent, Vincenzo De Luca, and James L Kennedy. Mixture regression analysis on age at onset in bipolar disorder patients: investigation of the role of serotonergic genes. *European Neuropsychopharmacology*, 20(9):663–670, 2010.

Charlotte H Mason and William D Perreault Jr. Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research*, 28(3): 268–280, 1991.

RH Mayers. Classical and modern regression with applications: Pwskent publ. *Co.: Boston*, 1990.

Gary C McDonald and Diane I Galarneau. A monte carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350): 407–416, 1975.

Geoffrey J McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

L Joseph Melton III, Elizabeth J Atkinson, W Michael O'Fallon, Heinz W Wahner, and B Lawrence Riggs. Long-term fracture prediction by bone mineral assessed at different skeletal sites. *Journal of Bone and Mineral Research*, 8(10):1227–1233, 1993.

L Joseph Melton III, Elizabeth J Atkinson, Michael K O'connor, W Michael O'fallon, and B Lawrence Riggs. Bone density and fracture risk in men. *Journal of Bone and Mineral Research*, 13(12):1915–1923, 1998.

Bart Meuleman, Geert Loosveldt, and Viktor Emonds. Regression analysis: Assumptions and diagnostics. *The SAGE handbook of regression analysis and causal inference*, pages 83–110, 2015.

Aitkin Murray. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55(1):117–128, 1999a.

Aitkin Murray. Meta-analysis by random effect modelling in generalized linear models. *Statistics in Medicine*, 18(17-18):2343–2351, 1999b.

W Navidi. Statistics for scientist and engineers, 2011.

Nelder and Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384, 1972.

Jenny Neuburger, Colin Currie, Robert Wakeman, Carmen Tsang, Fay Plant, Bianca De Stavola, David A Cromwell, and Jan van der Meulen. The impact of a national clinician-led audit initiative on care and mortality after hip fracture in england: an external evaluation using time trends in non-audit data. *Medical care*, 53(8):686–691, 2015.

Antony Ngunyi, Peter N Mwita, and Romanus O Otieno. On the estimation and properties of logistic regression parameters. 10(4):57–68, 2014.

Bruno Nicenboim, Shravan Vasishth, Felix Engelmann, and Katja Suckow. Exploratory and confirmatory analyses in sentence processing: A case study of number interference in german. *Cognitive Science*, 42:1075–1100, 2018.

Behdin Nowrouzi, Renan P Souza, Clement Zai, Takahiro Shinkai, Marcellino Monda, Jeffrey Lieberman, Jan Volvaka, Herbert Y Meltzer, James L Kennedy, and Vincenzo De Luca. Finite mixture regression model analysis on antipsychotics induced weight gain: Investigation of the role of the serotonergic genes. *European Neuropsychopharmacology*, 23(3):224–228, 2013.

Martin Paldam. Methods used in economic research: An empirical study of trends and levels. *Economics*, 15(1):28–42, 2021.

Sanjay Kumar Palei and Samir Kumar Das. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines: An approach. *Safety Science*, 47(1):88–96, 2009.

Andrew David Pearce and Armin Hatefi. Multiple observers ranked set samples for shrinkage estimators. *arXiv preprint arXiv:2110.07851*, 2021.

Raymond Pearl and Lowell J Reed. On the rate of growth of the population of the united states since 1790 and its mathematical representation. *Proceedings of the national academy of sciences*, 6(6):275–288, 1920.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

DAVID Peel and G MacLahlan. Finite mixture models. *John and Sons*, 2000.

Richard E Quandt and James B Ramsey. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 73(364):730–738, 1978.

Mamunur Rashid and Naima Shifa. Consistency of the maximum likelihood estimator in logistic regression model: A different approach. *Journal of Statistics*, 16 (1):1–11, 2009.

Phil Reed and Yaqionq Wu. Logistic regression for risk factor modelling in stuttering research. *Journal of fluency disorders*, 38(2):88–101, 2013.

Kathryn Roeder. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89(426): 487–495, 1994.

RL Schaefer, LD Roi, and RA Wolfe. A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, 13(1):99–113, 1984.

Peter Schlattmann. *Medical applications of finite mixture models*. Springer, 2009.

Nicholas J Schork, David B Allison, and Bonnie Thiel. Mixture distributions in human genetics research. *Statistical Methods in Medical Research*, 5(2):155–178, 1996.

Bo Segerstedt. On ordinary ridge regression in generalized linear models. *Communications in Statistics-Theory and Methods*, 21(8):2227–2246, 1992.

G Shmueli. To explain or to predict?. 25 (3), 289–310, 2010.

Ken R Smith, Martha L Slattery, and Thomas K French. Collinear nutrients and the risk of colon cancer. *Journal of clinical epidemiology*, 44(7):715–723, 1991.

Charles Stein. Variate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Contributions to the Theory of Statistics*, volume 1, page 197. University of California Press, 1956.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Steven M Trost and Garold D Oberlender. Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *Journal of construction Engineering and Management*, 129(2):198–204, 2003.

Jeanne Truett, Jerome Cornfield, and William Kannel. A multivariate analysis of

the risk of coronary heart disease in framingham. *Journal of chronic diseases*, 20 (7):511–524, 1967.

Yu-Kang Tu, Valerie Clerehugh, and Mark S Gilthorpe. Collinearity in linear regression is a serious problem in oral health research. *European Journal of Oral Sciences*, 112(5):389–397, 2004.

T Rolf Turner. Estimating the propagation rate of a viral infection of potato plants via mixtures of regressions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(3):371–384, 2000.

JC Van Houwelingen. Shrinkage and penalized likelihood as methods to improve predictive accuracy. *Statistica Neerlandica*, 55(1):17–34, 2001.

Peiming Wang and Martin L Puterman. Mixed logistic regression models. *Journal of Agricultural, Biological, and Environmental Statistics*, 3(2):175–200, 1998.

Michel Wedel, Frenkel Ter Hofstede, and Jan-Benedict EM Steenkamp. Mixture model analysis of complex samples. *Journal of Classification*, 15(2):225–244, 1998.

WHO. Assessment of fracture risk and its application to screening for postmenopausal osteoporosis: report of a who study group [meeting held in rome from 22 to 25 june 1992]. 1994.

Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Alternating minimization for mixed linear regression. In *International Conference on Machine Learning*, pages 613–621. PMLR, 2014.

Hong-Tu Zhu and Heping Zhang. Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66 (1):3–16, 2004.

Yajie Zou, Yunlong Zhang, and Dominique Lord. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention*, 50:1042–1051, 2013.