University of Nevada, Reno

Variable Selection in Linear Models with Grouped Variables

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy in
Statistics and Data Science

by

Jingxuan Yang

Mihye Ahn, Ph.D. Dissertation Advisor

Yinghan Chen, Ph.D. Dissertation Co-advisor

August, 2023

University of Nevada, Reno

THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

JINGXUAN YANG

entitled

Variable Selection in Linear Models with Grouped Variables

be accepted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

Mihye Ahn, Ph.D. *Advisor*

Yinghan Chen, Ph.D. *Co-advisor*

Deena Schmidt, Ph.D. *Committee Member*

Yan Liu, Ph.D. *Committee Member*

Teruni Lamberg, Ph.D. *Graduate School Representative*

Markus Kemmelmeier, Ph.D. Dean, *Graduate School*

August, 2023

Abstract

YANG, JINGXUAN. Variable Selection in Linear Models with Grouped Variables. (Under the direction of Dr. Mihye Ahn and Dr. Yinghan Chen)

Linear mixed models have been widely used for repeated measurements, longitudinal studies, or multilevel data. The selection of random effects in linear mixed models has received much attention recently in the literature. Random effects consider dependent structure between repeatedly measured data. Due to computational challenges, the selection of grouped random effects has yet to be studied. Grouped random effects, including genetics data or categorical variables, are commonly seen in practice. We present an efficient method for selecting random effects at group levels in linear mixed models. Specifically, the proposed method employs a restricted maximum likelihood function to estimate the covariance matrix of random effects. To achieve sparse estimation and grouped random effects selection, we then introduce a new shrinkage penalty term.

In addition, we extend the idea of grouped variable selection onto the latent regression model. By incorporating regression onto latent traits, latent regression models provide a way to uncover hidden influential factors from the data and make more accurate predictions. Specifically, we develop a variable selection approach for latent regression item response theory models by introducing the group LASSO penalty into the marginal log-likelihood function of observed test responses. We derive the explicit forms of updating steps for model parameters in a modified Newton-Raphson method. Our approach selects significant covariates and estimates model parameters simultaneously.

For both variable selection frameworks, we perform simulation studies to evaluate the variable selection performance of the proposed methods. We then compare them to existing or naive selection methods. Additionally, we apply the proposed methods on real data sets.

Dedication

This dissertation is dedicated to **my parents**, **Hui Yang** and **Quan Wang**, and **my child**, **Lucas**

For their love, support and encouragement

Acknowledgements

I am immensely thankful to my dissertation advisers. I want to thank Dr. Mihye Ahn, whose guidance and support have shaped my passion for Statistics and inspired me to delve deeper into my research topics. I also want to express my deep appreciation to my co-advisor, Dr. Yinghan Chen, whose willingness to advise me in a different research area has broadened my horizons and enriched my academic journey. I am truly grateful for their guidance, patience, and advice throughout my research and study life. I am fortunate to have had such exceptional mentors who have inspired and supported me along this academic journey.

I would like to extend my gratitude to the members of my Ph.D. committee, Dr. Deena Schmidt, Dr. Yan Liu, and Dr. Teruni Lamberg, for generously dedicating their time to evaluate my work.

I also really appreciate every faculty and staff member from the Department of Mathematics and Statistics at the University of Nevada, Reno. Their dedication and commitment to creating a friendly environment and well-designed curriculum have given students like me a chance to pursue career goals.

Lastly, I want to express my deep gratitude to my family and friends who have always supported and loved me in my life.

Contents

List of Tables

Chapter 1

Introduction

1.1 Background

In the field of Statistics, the topic of variable selection has received considerable interest in recent decades. Variable selection is a procedure that aims to find a set of the most useful variables in predicting a response variable. Many statisticians have developed efficient and accurate methodologies to select important variables to include in the model. Choosing relevant variables to add to the model helps in reducing the variance of parameter estimates. Additionally, having irrelevant variables in the model can cause the problem of overfitting, which means that the model fits the noise in the data. Overfitting usually results in poor predictive performance of the model. The procedure of selecting important variables can remove noise variables and alleviate the problem of overfitting, effectively improving the final model's predictive ability. Moreover, including many unnecessary variables can make interpreting the model much more challenging. By including only the most informative variables, statisticians can develop more parsimonious models that are easier to interpret. In practice, a simpler model can provide more meaningful insights into decision-making. Furthermore, modeling with the most important variables can reduce both time and

money costs and lead to more efficient modeling procedures.

A linear model, also called a linear regression model, is a statistical model that allows statisticians to study the linear relationship between a response variable and a set of independent variables or predictors. In linear regression problems, the primary goal is to estimate the regression coefficients in a manner that accurately characterizes the true relationship between the response variable and predictors. The standard estimation technique is the ordinary least squares (OLS) estimation, which finds the solution of regression coefficients by minimizing the sum of the squared errors. However, statisticians find OLS estimates to be unsatisfactory because they usually have a low bias but large variance, which can result in poor prediction accuracy. In addition, since the OLS method keeps all predictors in the final model, it does not give an easily interpretable model when there are many predictors. Hence, variable selection is considerably necessary in linear regression problems.

A wide range of variable selection methods in the linear model have been investigated in the literature. Best subset selection is one of the classical variable selection approaches. If there are p variables, the best subset selection considers all 2^p possible combinations of independent variables and selects the best subset of variables meeting some selection criteria. For example, if $p = 20$, the best subset selection selects the best subset of variables among $2^{20} = 1,048,576$ different combinations of candidate models. Thus, it can be computationally intense to utilize the best subset selection, especially when the number of variables p is very large. Other classical methods, including backward elimination, forward selection, and stepwise regression, can be employed to select variables with relatively low computational costs. Those methods choose variables sequentially without considering all possible combinations of predictors and stop picking when some specific selection criteria are met. Many well-defined selection criteria have been widely used in literature, including the Akaike informa-

tion criterion (AIC) (Akaike, 1973), Bayesian information criterion (BIC) (Schwarz, 1978) and Mallows' C_p (Mallows, 1973). In Section 1.2.1, we will provide a detailed explanation of several selection criteria. However, those classical variable selection methods may produce unstable selection results. Consequently, a slight change in data can lead to significant variability in the resulting models (Breiman, 1995).

1.1.1 Variable Selection in Linear Models

In recent years, there has been increasing literature on penalized likelihood methods in variable selection methods research. The general idea of the penalized methods is to estimate the regression coefficients by minimizing an objective function that contains two parts, a loss function and a penalty term. There are various choices for both loss functions and penalty terms. Several penalized likelihood methods have been developed in the recent literature, including non-negative garrote (Breiman, 1995), LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), Elastic net (Zou & Hastie, 2005), Adaptive LASSO (Zou, 2006; Zhang & Lu, 2007) and group LASSO (Yuan & Lin, 2006). Differently from other variable selection methods, group LASSO specifically focuses on selecting variables at the group level instead of the individual level in linear models. In Section 1.3, we will review these methods that have been designed to select and estimate regression coefficients in linear regression models.

1.1.2 Variable Selection in Linear Mixed Models

In linear models, the underlying assumption is that each observation is independent of each other. However, the assumption of independence might not hold true in all cases in real data. In order to account for the dependence structure in the data, Laird & Ware (1982) proposed a linear mixed model considering random effects in addition to fixed effects. This model is frequently employed for non-independent data, such

as longitudinal, spatial, panel, or multi-level data.

Variable selection on fixed and random effects are popular research topics in the literature. Researchers have developed various selection methods for selecting important fixed effects, random effects, or both in the framework of linear mixed models. Important fixed effects mean that the corresponding coefficients of those fixed-effect covariates are non-zero, and important random effects are the random effects whose coefficients have non-zero variance. Important random effects significantly contribute to the variability in the data. We will describe the rule of choosing random effects in detail in Section 1.4.

Selection of not only fixed effects but also random effects is crucial. In fact, the selection of random effects is closely related to the estimation of a variance-covariance matrix of random effects' coefficients. If significant random effects are omitted from the model, the covariance matrix would be underfitted and then could negatively affect the selection and estimation of fixed effects. Conversely, if insignificant random effects are wrongly added to the model, the covariance matrix may become nearly singular, leading to numerical instability for model fitting. Therefore, making a good selection of random effects can enhance the selection of fixed effects, consequently improving the prediction accuracy.

It has been a challenging problem to select important random effects in the linear mixed models due to the nature of their variance-covariance matrix. Many works of literature, including Bondell et al. (2010), Ahn et al. (2012), Pan & Shang (2018), Li et al. (2018), have studied this problem and given efficient ways to select significant random and fixed effects. We will present these existing methods in detail in Section 1.4.

We are motivated by the idea of group LASSO in the linear models and intend to select grouped random effects in the linear mixed models because no existing

methods have considered this case. We propose a novel method for the linear mixed models to select random effects at group levels. In addition, we extend the idea of grouped variable selection onto the latent regression models. Specifically, we develop a variable selection approach for latent regression item response theory models by introducing the group LASSO penalty into the marginal log-likelihood function of observed test responses.

1.1.3 Outlines of Dissertation

In Chapter 1, we introduce some background information and perform a literature review on existing variable selection methods in both linear models and linear mixed models. Chapter 2 showcases our first proposed method, which pertains to the random effects selection for grouped variables. We present our methodology, computational algorithm, and the process of tuning parameter selection in detail, from Sections 2.1 through 2.3. We then perform simulation studies and real data analysis, and draw comparisons with alternative methods in Section 2.4. Additionally, we follow a similar outline to describe our second variable selection methodology in latent variable models in Chapter 3. Furthermore, we include the extended comments on our current projects and possible future work in Chapter 4.

1.2 Classical Variable Selection Methods in Linear Models

Assume that the number of observations is n and the number of independent variables is p. In a matrix form, the linear regression model has the following model equation:

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

where $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ is the $n \times 1$ vector of response variables, $\boldsymbol{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is the $p \times 1$ vector of regression coefficients, and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^\top$ is the $n \times 1$ vector of random errors that is assumed to follow a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2 \boldsymbol{I}_n$. Because of the normality assumption of random errors $\boldsymbol{\epsilon}$, the response vector $\boldsymbol{y}$ is assumed to have a multivariate normal distribution with mean $\boldsymbol{X\beta}$ and variance-covariance matrix $\sigma^2 \boldsymbol{I}_n$.

The common approach to solving the linear regression problem is the ordinary least squares (OLS). The OLS estimator has an explicit form as follows:

$$\widehat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{1.1}$$

Unfortunately, the OLS estimates include all the ordinary variables in the model. It means that it is not able to perform variable selection. Therefore, to obtain a more interpretable model, some variable selection techniques are necessary to be employed if there are many independent variables.

In the literature, many selection criteria and variable selection methods have been proposed to identify the most informative set of predictors for inclusion in the model, with the goal of ultimately obtaining the best model. In the following subsection, we will review several selection criteria that have been commonly used in practice.

1.2.1 Selection Criteria

1.2.1.1 Mean Squared Error (MSE)

The mean square error (MSE) is defined as

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{n - p} = \frac{RSS}{n - p},$$

where $\widehat{y}_i$ is the fitted value of y_i, n is the number of observations, and p is the number of predictors in the fitted model. RSS represents the residual sum of squares, measuring the discrepancy between the estimated model and the actual data. The model with the smallest MSE is preferred among all candidate models. In the case of small data sets, the effectiveness of MSE may be limited.

1.2.1.2 Coefficient of Determination: R^2

The coefficient of determination, R^2, is defined as

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{RSS}{SST},$$

where $\bar{y}$ is the mean of y, SSR is the regression sum of squares, and SST is the total sum of squares. The coefficient of determination measures the proportion of variance in the response variable that the predictors can explain. It assesses how well the data fit the model and takes a value between 0 and 1. The selection rule of using R^2 is to choose the model with the largest R^2. However, since R^2 always increases when more predictors enter the model, model selection based on R^2 might result in the problem of overfitting.

1.2.1.3 Adjusted R^2

To overcome the drawback of R^2, a modified version of R^2, called adjusted R^2, is proposed. The adjusted R^2 has the form as follows:

$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{SST/(n-1)} = 1 - \frac{(n-1)MSE}{SST} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

Similar to the rule using R^2, we choose the model with the largest adjusted R^2. However, the adjusted R^2 increases only when the inclusion of additional predictors

improves the overall model fit. Additionally, it is straightforward to see that minimizing MSE is equivalent to maximizing adjusted R^2. Therefore, comparing models in terms of MSE or adjusted R^2 can give the same selection result.

1.2.1.4 AIC

The concept of the Akaike information criterion (AIC), first introduced by Akaike (1973), offers a way to compare various candidate models and select the best model among them. It is one of the most popular selection tools nowadays. The AIC incorporates a penalty for model complexity to the log-likelihood of the model so that it penalizes models that use more parameters. The general form of AIC is defined as

$$AIC = -2\log(likelihood) + 2p.$$

However, AIC might select overly complex models in the small sample, leading to overfitting problems. Then, an adjusted version of AIC was proposed by Hurvich & Tsai (1989), called corrected AIC. The corrected AIC has the following form

$$AIC_c = AIC + \frac{2(p+1)(p+2)}{n-p-2}.$$

The corrected AIC improves the model selection performance in a small sample setting by considering the sample size in the penalty. It induces a heavier penalty on the number of parameters. Among a set of candidate models, the one with the smallest AIC or corrected AIC is preferred. Several other variants of AIC have also been studied to solve various problems (McQuarrie & Tsai, 1998).

1.2.1.5 BIC

Bayesian Information Criterion (BIC) is another popular information criterion in model selection, proposed by Schwarz (1978). The idea of BIC is derived from Bayesian principles, and it is defined as

$$BIC = -2\log(likelihood) + p\log n.$$

Both BIC and AIC are likelihood-based methods. The only difference between calculating BIC and AIC is the multiplier of p in the penalty term. Consequently, when $n > e^2$, BIC penalizes more on the number of parameters. Therefore, BIC tends to select more parsimonious models. Additionally, BIC exhibits asymptotic consistency, implying that as the sample size approaches infinity, the probability of choosing the correct model using the BIC approaches 1. In other words, as the sample size grows infinitely large, BIC can always select the true model if the true model happens to be under consideration. The consistency property of BIC is attractive for statisticians to use in practice. However, BIC has some limitations as well. Since BIC penalizes more complex models, it tends to select overly simple models in finite samples, which may result in underfitting. Moreover, Hurvich & Tsai (1989) demonstrated that BIC might have poor selection performance in small samples.

1.2.1.6 Mallows' C_p

The statistic known as Mallows' C_p (Mallows, 1973) is used to compare models that have different subsets of parameters when compared to the full model. It can be calculated using the following form:

$$C_p = \frac{RSS_p}{\widehat{\sigma}^2} - n + 2p,$$

where RSS_p is to the RSS of the reduced model, while $\widehat{\sigma}^2$ is the MSE of the full model. According to Murtaugh (1998), a stepwise procedure is employed to add or remove predictors until the smallest value of C_p is achieved. To compare models using this statistic, selecting the model that minimizes the C_p criterion and $C_p \approx p$ is recommended.

1.2.1.7 PRESS

The prediction sum of squares (PRESS) (Allen, 1974) is a statistic that can assess a model's predictive ability. The formula of PRESS is given by

$$PRESS = \sum_{i=1}^{n} (y_i - \widehat{y}_{(i)})^2,$$

where $\widehat{y}_{(i)}$ refers to the predicted value of the ith observation y_i obtained from the model that is fitted without including y_i. In other words, to calculate PRESS, each observation is excluded one at a time, and a linear regression model is fitted to the remaining $n-1$ observations to predict the value of the omitted response variable. Smaller PRESS indicates the better predictive ability of the model. Therefore, the model with the minimum PRESS is preferred.

PRESS utilizes all data and avoids data-splitting difficulties to validate the models (Holiday et al., 1995). On the other hand, since the calculation of PRESS requires fitting models n times, it is highly time-consuming to compute PRESS when the sample size n is large. Moreover, Breiman & Spector (1992) demonstrated that non-resampling estimates, such as the PRESS, can result in imprecise estimates of the mean squared error of prediction. They proposed using resampling techniques, such as cross-validation and bootstrap methods, to address this issue.

1.2.1.8 Cross Validation

Cross-validation is a model evaluation method that measures the model's predictive performance on unseen data. In cross-validation, the data is usually divided into two subsets. One subset, called training data, is used to train the model; the other subset, called test data, is used to test its performance. Among various types of cross-validation techniques, K-fold cross-validation is one of the most widely used methods. Specifically, the K-fold cross-validation randomly splits the sample data into K equal-sized groups. Then, the model is trained using $K-1$ groups and is tested using the remaining group. The process repeats K times until each group gets a chance to be the test group. The model's performance can be averaged over K repetitions to give an overall estimation. In practice, 5 and 10 are commonly used choices for K.

Leave-one-out cross-validation (LOOCV) is a special case of K-fold cross-validation where K equals n. In other words, each observation is considered as a separate group. PRESS statistic uses the idea of LOOCV in its computation. However, when the sample size is large, it is considerably computationally expensive to perform LOOCV. In this case, K-fold cross-validation with $K \ll n$ might be more appropriate.

When selecting a model, the one with the lowest squared error from cross-validation is usually considered to be the best option. We then pick the model with the lowest MSE, averaged across testing sets.

1.2.1.9 Generalized Cross Validation (GCV)

Generalized Cross-Validation (GCV) method is a computational shortcut of LOOCV (Hastie et al., 2009), proposed by Craven & Wahba (1979). It attempts to reduce the computational burden of cross-validation and provide an approximation to LOOCV.

The general form of GCV approximation is defined as

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y}_i}{1 - \mathrm{tr}(\boldsymbol{S})/n}\right)^2,$$

where $\boldsymbol{S}$ is the matrix for $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ and $\mathrm{tr}(\boldsymbol{S})$ refers to the effective number of degrees of freedom.

In linear models, $\mathrm{tr}(\boldsymbol{S}) = p$. Thus, GCV can be written as follows

$$GCV = \frac{RSS}{n} \cdot \frac{1}{(1 - p/n)^2}.$$

In practice, it is challenging to calculate each diagonal element in the matrix $\boldsymbol{S}$, which motivates statisticians to propose various GCV-type statistics to address this issue. Several different GCV-type statistics will be described and utilized in the proposed method in Section 3.4.

1.2.2 Computational Techniques

When there are many predictors in the data set, it is necessary to choose a subset of predictors significantly related to the response variable and then include this subset of predictors to build the best model. Many statisticians have proposed various methods to perform variable selection to improve the selection accuracy and computation efficiency. This section will describe several classical variable selection methods widely used in practice.

1.2.2.1 Best Subset Selection

Best subset selection involves considering every possible combination of the potential predictors. Specifically, if there are p predictors, 2^p candidate models will be considered. The best model is determined by certain selection criteria through the

process of best subset selection from all the candidate models. For example, based on adjusted R^2, the candidate model with the largest R^2_{adj} is selected, and the predictors in this model are considered the best independent variables. Besides adjusted R^2, many other selection criteria described in Section 1.2.1, such as Mallows' C_p, MSE, *etc.*, can also be used in the best subset selection. Even though the best subset selection is easy to understand and implement, in practice, it requires a significantly massive computation when the number of predictors is large. Additionally, best subset selection tends to select overly complex models potentially leading to overfitting problems.

Since best subset selection requires expensive computation, statisticians usually prefer alternative methods to select variables. In the following, we will present several traditional variable selection methods that are more computationally efficient.

1.2.2.2 Backward Elimination

Backward elimination is a straightforward and efficient model selection approach. The process of backward elimination requires at most $1 + p(p+1)/2$ model fittings to identify the best model, which is significantly lower than 2^p required in best subset selection as p grows. The general idea of backward elimination is that it gradually removes unimportant variables until only important variables are left in the model. Specifically, it starts with considering a full model that includes all potential predictors from the data. The next step drops the least statistically significant variable one at a time. In the traditional implementation of backward elimination, the significance of each variable is assessed using the F-statistic. The variable with the smallest F-statistic, which indicates the least significant variable, is then deleted from the model. This process is repeated until the remaining variables are all significant at a pre-determined significance level. Besides F-statistic, other selection

criteria, such as adjusted R^2, BIC, AIC, *etc.*, can also be implemented at each step as a criterion to decide the deletion of a predictor. If the value of the selected criterion is not improved during the procedure, backward elimination stops.

It is worth noting that backward elimination can only be employed when the number of predictors is less than the number of observations, as the method starts with the full model that includes every variable in the model. Thus, backward elimination is not a good choice for variable selection in a high-dimensional setting. Moreover, in backward elimination, once the predictor is removed, it is not re-entered to the model even if it becomes significant later in model fitting.

1.2.2.3 Forward Selection

Forward selection works in the opposite direction of backward elimination. Similar to backward elimination, forward selection also needs to fit at most $1 + p(p+1)/2$ models to obtain the best model. However, instead of starting with a full model, forward selection starts with a null model that includes no variables in the model. The next step is to add one variable at a time to the model based on its statistical significance. Among the variables that are not included in the model, the one with the largest F-statistic is added to the model. Other pre-specified criteria, such as AIC or BIC, can also be implemented to decide if the variable should be added or not. The above step is repeated until every variable in the model is significant based on F-statistic, or the selected criterion is no longer improved by adding any of the remaining variables.

Forward selection is useful when the number of predictors from data is large because it reduces the computational complexity of searching for the best model by starting with a null model. In a study by Roecker (1991), using forward selection can result in a slight reduction in prediction error and bias when compared to using all

possible regression models. Different from backward elimination, once the predictor is added to the model, it always stays in the model.

1.2.2.4 Stepwise Selection

Stepwise regression is a combination of backward elimination and forward selection. Its selection procedure can either go backward (start with a full model) or forward (begin with a null model). At each step, the procedure either enters or removes one variable at a time based on the pre-specified selection criterion. The main difference in stepwise selection is that a variable that has been removed in a previous step can be added back to the model if it is found to be significant later on.

Those classical methods presented above are easy to implement in practice. However, their selection results are not stable because of inherent discreteness (Breiman, 1995; Fan & Li, 2001). A small change in the data may result in quite different variable selection results. Moreover, the unstable performance of those methods might lead to worse prediction accuracy. In the next section, we will review several penalized likelihood methods that provide more reliable selection results than the stepwise methods do.

1.3 Penalized Likelihood Methods in Linear Models

Penalized likelihood methods, also called shrinkage methods, work as variable selection approaches by adding a particular type of penalty term to the likelihood function. This penalty term is designed to shrink some regression coefficients towards zeros, while shrinking some small coefficients to exactly zeros. By such shrinkage on regression coefficients, many unimportant variables are excluded from the model. One

advantage of penalized likelihood methods is that they can effectively avoid overfitting by penalizing the regression coefficients of insignificant variables. Additionally, it improves the interpretability of the model by only keeping the most informative variables in the model. Furthermore, some methods are able to estimate regression coefficients and select variables simultaneously.

The general form of penalized likelihood approaches is to minimize the following objective function

$$L(\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}), \tag{1.2}$$

where $\boldsymbol{\beta}$ represents a vector of regression coefficients, $L(\boldsymbol{\beta})$ represents a loss function, such as the negative likelihood function or least square loss function among other various options, $P(\boldsymbol{\beta})$ is a penalty function, and λ is a tuning parameter, which is a non-negative constant. λ controls the strength of penalization and thus controls the complexity of the model. When $\lambda = 0$, the second term in (1.2) disappears, which means that there is no penalty added. For example, if the loss function is the least square loss function and $\lambda = 0$, no variables are eliminated, and the problem is equivalent to ordinary least squares estimation. As the value of λ increases, the penalty term also increases, resulting in greater shrinkage imposed on the coefficients. When λ is sufficiently large, all regression coefficients can be shrunk to zero. Moreover, the value of λ is closely related to the bias and variance of the model. As the value of λ increases, the bias in the model also increases, whereas decreasing λ leads to a higher variance. For instance, choosing a small λ results in a bigger model with a lower bias, but it comes with the trade-off of a much larger variance. Thus, selecting an appropriate value for λ is crucial to achieving accurate and efficient variable selection. Using the criteria for variable selection presented in Section 1.2.1, we can determine the appropriate value of λ.

Many penalized likelihood methods have been developed by statisticians, consid-

ering various likelihood functions and penalty terms. In the following subsections, we will review several penalized likelihood approaches.

1.3.1 Ridge Regression

Ridge regression is the first penalized regression method proposed by Hoerl & Kennard (1970b,a). The penalty term in ridge regression is L_2 penalty, also called L_2 norm, defined as the sum of the squared coefficients. The ridge estimate is defined as

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right), \tag{1.3}$$

with $\lambda \geq 0$. We assume both design matrix $\boldsymbol{X}$ and responses $\boldsymbol{y}$ are centered. Thus, the intercept term is ignored in (1.3). The ridge estimator is given by

$$\widehat{\boldsymbol{\beta}}_{ridge} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y},$$

where $\boldsymbol{I}$ is a $p \times p$ identity matrix. It is easy to see that the ridge estimator is analogous to OLS estimator in (1.1). Even though $\widehat{\boldsymbol{\beta}}_{OLS}$ is an unbiased estimator of $\boldsymbol{\beta}$, it does not perform well when $\boldsymbol{X}$ is ill-conditioned. Since $\widehat{\boldsymbol{\beta}}_{OLS}$ involves the term $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, it is infeasible to compute $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ if $\boldsymbol{X}^\top \boldsymbol{X}$ is singular or nearly singular. In this case, $\boldsymbol{X}$ is called ill-conditioned because a small change in the elements of $\boldsymbol{X}$ could result in a large change in $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. The extra term $\lambda \boldsymbol{I}$ in the ridge estimator makes it a biased estimator but also improves the coefficient estimation in the case of ill-conditioned $\boldsymbol{X}$.

Ridge regression successfully shrinks the estimates towards zero and produces an improved estimation than OLS. Compared with stepwise selection, ridge regression is more stable in selecting variables, but it does not shrink any regression coefficients to exact zero. Thus, it cannot be used as a tool for variable selection.

1.3.2 Non-negative Garrote

Breiman (1995) presented an innovative method, called non-negative garrote, to select the best subset of variables. Let us assume that $\boldsymbol{X}$ is standardized, and the response vector $\boldsymbol{y}$ has a mean of zero. Then, the non-negative garrote intends to find a set of non-negative slicing factors c_j to minimize

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}c_j\widehat{\beta}_j^{OLS}x_{ij}\right)^2+\lambda\sum_{j=1}^{p}c_j \qquad \text{subject to} \quad c_j\geq 0,$$

where $\widehat{\beta}_j^{OLS}$ is the OLS estimates and $\lambda \geq 0$. The garrote estimates are $\widehat{\beta}_j^{garrote} = c_j\widehat{\beta}_j^{OLS}$ for $j = 1,\ldots,p$. As λ increases, more c_j shrinks to exact zero. Therefore, the non-negative garrote is able to produce sparse models. Additionally, Zou (2006) proved that variable selection by the non-negative garrote is consistent.

1.3.3 LASSO

Tibshirani (1996) introduced the least absolute shrinkage and selection operator (LASSO) as a new regression method. The objective function includes the least squares loss function and L_1 penalty. The L_1 penalty, also called L_1 norm, takes the sum of the absolute value of the magnitude of coefficients. The LASSO estimator is defined by

$$\widehat{\boldsymbol{\beta}}_{LASSO}=\underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right),$$

where $|\beta_j|$ is the absolute value of β_j and $\lambda \geq 0$ is a tuning parameter. Due to the property of L_1 penalty, LASSO is able to shrink some coefficients to exact zero. Therefore, LASSO is capable of selecting informative variables and estimating regression coefficients efficiently at the same time.

Leng et al. (2006) found that the LASSO estimate does not give consistent model

selection in a setting of the fixed p and orthogonal designs. Zou (2006) examined a necessary condition for the LASSO to be consistent, whereas Zhao & Yu (2006) presented an almost necessary and sufficient condition for a consistent LASSO solution in the fixed p setting and in the large p setting as the sample size gets larger. Meinshausen & Bühlmann (2006) also demonstrated that, under some conditions, LASSO can provide consistent estimates of the dependency between Gaussian variables in high-dimensional settings. Therefore, the LASSO may not always be consistent.

LASSO has certain limitations as well. Firstly, if the number of predictors exceeds the number of observations, indicated by $p > n$, then LASSO selects at most n predictors before it reaches saturation. Secondly, when some predictors are highly correlated with each other, LASSO tends to choose one of those predictors and does not consider which one is selected. Lastly, for the case where $n > p$ with collinearity predictors, empirical studies suggest that ridge regression outperforms LASSO in terms of predictive performance (Tibshirani, 1996).

1.3.4 Elastic Net

Motivated by the limitations of LASSO, Zou & Hastie (2005) proposed an elastic net that takes advantage of both LASSO and ridge regression. The elastic net estimate is defined by

$$\widehat{\boldsymbol{\beta}}_{elastic\ net} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right), \tag{1.4}$$

where both $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters. The penalty term in (1.4) combines L_1 and L_2 penalties. Elastic net does a good job of selecting variables for high-dimensional data, where $p >> n$.

1.3.5 SCAD

Fan & Li (2001) proposed a novel penalty function, smoothly clipped absolute deviation (SCAD). The SCAD penalty imposes different penalty terms on the regression coefficients based on the relationship between the magnitude of coefficients and the tuning parameter λ. For each regression coefficient β_j, the SCAD penalty is given by

$$p_{SCAD}(\beta_j) = \begin{cases} \lambda|\beta_j|, & \text{for } |\beta_j| \leq \lambda \\ -(\dfrac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}), & \text{for } \lambda < |\beta_j| \leq a\lambda \\ \dfrac{(a+1)\lambda^2}{2}, & \text{for } |\beta_j| > a\lambda, \end{cases}$$

where $a > 2$ and $\lambda > 0$.

Then, the SCAD estimates can be defined by

$$\widehat{\boldsymbol{\beta}}_{SCAD} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \sum_{j=1}^{p} p_{SCAD}(|\beta_j|) \right).$$

Additionally, Fan & Li (2001) introduced 'oracle properties' of a variable selection procedure. In their opinion, a good variable selection procedure should have such properties. Let $\widehat{\boldsymbol{\beta}}(\gamma)$ denote the coefficient estimator produced by a procedure γ, and A denote the set of non-zero regression coefficients in the true model. Then, an oracle procedure γ is called if $\widehat{\boldsymbol{\beta}}(\gamma)$ has the oracle properties that:

1. identifies the correct model, *i.e.* $\{j : \widehat{\beta}_j \neq 0\} = A$, for $j = 1, \ldots, p$.
2. has the optimal estimation rate, $\sqrt{n}(\widehat{\boldsymbol{\beta}}(\gamma)_A - \boldsymbol{\beta}^*_A) \to N(\mathbf{0}, \boldsymbol{\Sigma}^*)$ in distribution, where $\boldsymbol{\beta}^*$ refers to the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}^*$ refers to the variance-covariance matrix of the true predictors.

They have shown that the SCAD penalty exhibits oracle properties, which means that it is capable of consistent selection and optimal estimation asymptotically. Con-

versely, the LASSO penalty lacks these properties. However, since the SCAD penalty is a non-concave function and non-differentiable at zero, there is no guarantee that the local maximum of the penalized likelihood is the global maximum. This property makes it challenging to find the optimal solution.

1.3.6 Adaptive LASSO

Zou (2006) showed that LASSO can be consistent under a necessary condition. He also mentioned some scenarios in which the LASSO selection cannot be consistent. He then proposed a modification of LASSO, called adaptive LASSO. Assume $\widetilde{\boldsymbol{\beta}}$ is a root-n-consistent estimator to $\boldsymbol{\beta}^*$. The solution of adaptive LASSO is shown as follows:

$$\widehat{\boldsymbol{\beta}}_{ALASSO} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \widehat{\omega}_j |\beta_j| \right),$$

where $\lambda \geq 0$ and the weight $\widehat{\omega}_j = 1/|\widetilde{\beta}_j|^{\gamma_j}$ for $\gamma_j > 0$.

Compared to the LASSO penalty, the penalty term of adaptive LASSO contains an additional weight in front of the magnitude of regression coefficients. Zou (2006) showed that adaptive LASSO has the oracle properties.

1.3.7 Group LASSO

Yuan & Lin (2006) extended the idea of LASSO to solve the problem of selecting group variables in order to achieve better prediction accuracy in regression problems. Group variables are commonly seen in regression problems: one example is that a multi-level categorical variable is usually represented in a group of dummy variables; another example is the additive model with continuous variables, where a continuous variable could be represented in a linear combination of some basis functions of the measured variable. Both examples contain group structure, and the group LASSO

intends to select those important group variables rather than individual predictors.

Let us explore the concept of group variables using a real data example. For this purpose, we can analyze the birth weight data set presented in Hosmer Jr & Lemeshow (1989). This birth weight data set contains the birth weights of 189 newborns and several predictors containing the mother's information. In the following two examples, we will specifically focus on two predictors: a categorical variable representing the mother's race, which consists of three levels (*white*, *black*, or *other*), and a continuous variable indicating the mother's weight in pounds at the last menstrual period. Denote the newborn's birth weight as $\boldsymbol{Y}$, mother's race as $\boldsymbol{X}_1$, and mother's weight as $\boldsymbol{X}_2$.

Example 1: If we are interested in the relationship between the newborn's birth weight ($\boldsymbol{Y}$) and the mother's race ($\boldsymbol{X_1}$), we can fit a one-way analysis of variance (ANOVA) model, which is given by:

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_{1\cdot Balck} + \beta_2 \boldsymbol{X}_{1\cdot Other} + \boldsymbol{\epsilon},$$

where β_0 is the intercept, β_1 and β_2 are regression coefficients, $\boldsymbol{X}_{1\cdot Balck}$ and $\boldsymbol{X}_{1\cdot Other}$ are dummy variables of *black* and *other* levels, respectively, and $\boldsymbol{\epsilon}$ are the random errors. Here, the reference level is *white*. To select important variables, the group LASSO approach treats two dummy variables, $\boldsymbol{X}_{1\cdot Balck}$ and $\boldsymbol{X}_{1\cdot Other}$, of the predictor *mother's race* as a group. While the LASSO may choose either $\boldsymbol{X}_{1\cdot Balck}$ or $\boldsymbol{X}_{1\cdot Other}$ or both, the group LASSO selects or eliminates the entire group, treating all levels in *race* as a cohesive unit.

Example 2: If a non-linear effect of the mother's weight ($\boldsymbol{X}_2$) on the newborn's weight ($\boldsymbol{Y}$) exists, a polynomial additive model can be considered. Suppose we

employ a third-order polynomial regression on $\boldsymbol{Y}$:

$$\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_2 + \beta_2 \boldsymbol{X}_2^2 + \beta_3 \boldsymbol{X}_2^3 + \boldsymbol{\epsilon},$$

where β_0 is the intercept, β_1,β_2 and β_3 are regression coefficients, $\boldsymbol{X}_2^2$ represents the quadratic term of mother's weight, and $\boldsymbol{X}_2^3$ represents the cubic term of mother's weight. It is important to note that in this scenario, the group LASSO approach treats $\boldsymbol{X}_2$ and its quadratic and cubic terms as a group, as they are all derived from the same variable $\boldsymbol{X}_2$. Therefore, the group LASSO would either select or exclude $\boldsymbol{X}_2$, $\boldsymbol{X}_2^2$ and $\boldsymbol{X}_2^3$ simultaneously as a cohesive group, while the LASSO would likely select only one or two of these terms individually. From the above two examples of group variables, we can see that group variables are commonly seen in practice in regression problems.

The linear regression model can be written in terms of m groups of variables:

$$\boldsymbol{y} = \sum_{l=1}^{m} \boldsymbol{X}^{(l)} \boldsymbol{\beta}^{(l)} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, $\boldsymbol{X}^{(l)}$ is an $n \times p_l$ design matrix corresponding to the lth group of predictors, and $\boldsymbol{\beta}^{(l)}$ is the vector of regression coefficients in the lth group with group size p_l for $l = 1, \ldots, m$. Let p denote the total number of predictors, that is, $p = \sum_{l=1}^{m} p_l$. The group LASSO estimator is defined as:

$$\widehat{\boldsymbol{\beta}}_{GLASSO} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left|\left| \boldsymbol{y} - \sum_{l=1}^{m} \boldsymbol{X}^{(l)} \boldsymbol{\beta}^{(l)} \right|\right|_2^2 + \lambda \sum_{l=1}^{m} \sqrt{p_l} ||\boldsymbol{\beta}^{(l)}||_2,$$

where $\lambda \geq 0$ is a tuning parameter.

Interestingly, when $p_1 = \cdots = p_m = 1$, such as only one individual in each group, the group LASSO reduces to the LASSO. Moreover, when $m = 1$, that means there only exists one group with p variables, the group LASSO becomes equivalent to ridge

regression. The idea of group LASSO has also been extended to the logistic regression model (Lukas et al., 2008).

The group LASSO has gained significant popularity in practice due to its advantages over the LASSO:

1. Group variable selection: the group LASSO performs group variable selection instead of individual variable selection. It is particularly advantageous in scenarios such as ANOVA models and additive models with polynomial terms, where selecting groups of variables is often more desirable. For example, when dealing with categorical variables, the LASSO may select individual dummy variables instead of choosing whole factors (Lukas et al., 2008). However, in practice, it is generally preferred to select all levels within a categorical variable together rather than separately.

2. Robustness to dummy variable encoding: the group LASSO is more robust to the encoding of dummy variables compared to the LASSO. The variable selection results from the LASSO can be affected by how dummy variables of categorical variables are encoded in modeling. Different ways of encoding dummy variables may result in different selection outcomes. Since the group LASSO selects variables at the group level, it is much less sensitive to encoding of dummy variables.

3. Enhanced Interpretability: by selecting entire groups of variables together, the group LASSO significantly improves the interpretability of the final model.

4. Improved prediction accuracy: by incorporating group structures, the group LASSO can improve prediction accuracy in some instances. The group structure helps to detect important interactions and dependencies among variables.

1.4 Variable Selection in Linear Mixed Models

As mentioned in Section 1.1, the linear mixed model incorporates both fixed and random effects. Assume that there are m subjects (or clusters), and subject i gets n_i measurements. Let $N = \sum_{i=1}^{m} n_i$ be the total number of observations. A linear mixed model has a general form for each subject i:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, ..., m, \tag{1.5}$$

where $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})^\top$ is an $n_i \times 1$ vector of responses for subject i, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}^\top, ..., \boldsymbol{X}_{in_i}^\top)^\top$ is an $n_i \times p$ matrix of fixed-effect covariates for subject i, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ is a $p \times 1$ vector of fixed-effect coefficients, $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}^\top, ..., \boldsymbol{Z}_{in_i}^\top)^\top$ is an $n_i \times q$ matrix of random-effect covariates for subject i, $\boldsymbol{b}_i = (b_{i1}, ..., b_{iq})^\top$ is a $q \times 1$ vector of random-effect coefficients, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, ..., \epsilon_{in_i})^\top$ is an $n_i \times 1$ random error vector.

Similar to the linear model, the linear mixed model also has a few assumptions: 1) Each observed individual is independent to each other. 2) There exists linearity in fixed and random effects covariates, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. 3) The random effects coefficients $\boldsymbol{b}_i$ follows a multivariate normal distribution, that is $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{D})$, where the variance-covariance matrix $\boldsymbol{D}$ is a symmetric and positive semi-definite. 4) The residuals $\boldsymbol{\epsilon}_i$'s are independent and identically distributed as a multivariate normal distribution, that is $\boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}_{n_i})$, where $\boldsymbol{I}_{n_i}$ is an $n_i \times n_i$ identity matrix with 1's on the diagonal of the matrix and 0's are off-diagonal. 5) $\boldsymbol{b}_i$ and fixed effects covariates $\boldsymbol{X}_i$'s are independent. Based on those assumptions, it can be derived that

$$\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta},\ \sigma^2\boldsymbol{V}_i),$$

where $\boldsymbol{V}_i = \boldsymbol{I}_{n_i} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\top$. Therefore, the variance-covariance matrix $\boldsymbol{D}$ is closely

related to selecting random effects.

Considerable research has been devoted to variable selection for both fixed and random effects. The literature encompasses studies on various information criteria as well as investigations into the effectiveness of penalized likelihood approaches for selecting either fixed effects, random effects, or both. In this section, we will review several information criteria and penalized likelihood methods in the framework of linear mixed models.

1.4.1 Information Criteria

Information criteria are widely adopted in model selection in linear mixed models. The AIC is one of the most frequently utilized information criteria in model selection. While the AIC has been frequently used in the linear models, it has also been constructed within the framework of the linear mixed model. For example, the marginal AIC ($mAIC$) is a commonly used AIC in the linear mixed models. The $mAIC$ was developed based on the marginal likelihood of responses, which is defined as

$$mAIC = -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}^2) + 2a_N(p+q),$$

where $\ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}^2)$ is the maximized likelihood with the maximum likelihood estimates (MLE) or restricted maximum likelihood estimates (REML) of parameters, for the fixed effects $\widehat{\boldsymbol{\beta}}$ and for variance-covariance of random effects $\widehat{\boldsymbol{\theta}}$. $a_N = 1$ in the infinite sample form and $a_N = N/(N-p-q-1)$ in the finite sample form (Sugiura, 1978). Additionally, p is the number of parameters in the fixed effects and q is the number of parameters in the variance-covariance matrix $\boldsymbol{V}_i$. However, it has been shown that $mAIC$ is positively biased for the marginal Akaike Information, and there is no simple bias correction to make $mAIC$ exactly unbiased (Greven & Kneib, 2010).

Vaida & Blanchard (2005) derived the AIC based on the conditional model formulation, called conditional AIC ($cAIC$). It is formulated as

$$cAIC = -2\log[f(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}, \widehat{\boldsymbol{\theta}})] + 2K,$$

where $f(\cdot)$ is the conditional likelihood and K is the effective number of degrees of freedom.

Another widely used information criterion is BIC. The simplest BIC in linear mixed models is obtained by replacing $2a_N$ in the penalty term by $\log(N)$ in $mAIC$. It is called the marginal BIC ($mBIC$), which is given by:

$$mBIC = -2\ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}^2) + \log(N)(p+q).$$

The increased weight on the term $(p+q)$ encourages $mBIC$ to pick smaller models than $mAIC$ does.

1.4.2 Penalized Likelihood Methods

Penalized likelihood methods have been successfully extended to the linear mixed models. The selection of random effects is one of the challenges for linear mixed models. To remove an unimportant random effect, an entire row and column of $\boldsymbol{D}$ should be eliminated. Therefore, using penalized methods properly on random effects is challenging. Recently, many penalized variable selection methods on linear mixed models have been proposed.

Bondell et al. (2010) proposed a penalized joint variable selection method for both fixed and random effects in linear mixed models. They adopted a modified Cholesky decomposition on the covariance matrix of random effects $\boldsymbol{D}$ from Z. Chen & Dunson (2003), such that

$$\boldsymbol{D} = \widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{\Gamma}}^{\top}\widetilde{\boldsymbol{D}},$$

where $\widetilde{\boldsymbol{D}}$ is a diagonal matrix with $d_1, d_2, \ldots, d_q$ and $\widetilde{\boldsymbol{\Gamma}}$ is a lower triangular matrix with 1's on the diagonal. They also added adaptive penalty terms on the log-likelihood function to select fixed and random effects, respectively. According to their reparameterization, the linear mixed model can be written as:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\widetilde{\boldsymbol{D}}\widetilde{\boldsymbol{\Gamma}}\widetilde{\boldsymbol{b}}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{y}_i$ is a centered response for the ith subject, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are design matrices for fixed and random effects, respectively, $\widetilde{\boldsymbol{b}}_i = (b_{i1}, \ldots, b_{iq})^{\top}$ is a $q \times 1$ coefficient vector for random effects, and $\boldsymbol{\epsilon}_i$ is a random error vector. They denote $\boldsymbol{\phi} = (\boldsymbol{\beta}^{\top}, \boldsymbol{d}^{\top}, \boldsymbol{\gamma}^{\top})^{\top}$ as a parameter set, where $\boldsymbol{d} = (d_1, \ldots, d_q)^{\top}$ is a vector of the diagonal elements of $\widetilde{\boldsymbol{D}}$ and $\boldsymbol{\gamma}$ is a vector of the $q(q-1)/2$ free elements of $\widetilde{\boldsymbol{\Gamma}}$.

After reparameterization and treating random effects coefficients as observed, they derived the complete data log-likelihood function:

$$L_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b}) = -\frac{N + mq}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(||\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{D}^*\boldsymbol{\gamma}^*\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \boldsymbol{b}^{'}\boldsymbol{b}),$$

where $\boldsymbol{Z}$ is a block diagonal matrix of $\boldsymbol{Z}_i$, $\boldsymbol{D}^* = I_m \otimes \widetilde{\boldsymbol{D}}$ and $\boldsymbol{\Gamma}^* = I_m \otimes \widetilde{\boldsymbol{\Gamma}}$, with $\otimes$ represents the Kronecker product.

Since $||\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{D}^*\boldsymbol{\gamma}^*\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||$ is the only term that related to fixed effects and random effects in the complete data log-likelihood, they define the objective function by adding the L_1 penalty with the adaptive weights to this norm term:

$$\boldsymbol{Q}_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b}) = ||\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{D}^*\boldsymbol{\gamma}^*\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_m\left(\sum_{j=1}^{p}\frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q}\frac{|d_j|}{|\bar{d}_j|}\right),$$

where $\bar{\beta}_j$ represents the generalized least squares estimate of β_j, and $\bar{d}_j$ represents the decomposed component of the estimated covariance matrix obtained from the restricted maximum likelihood. Here, the penalty terms of fixed and random effects use the same λ_m as the tuning parameter. They implemented the constraint EM algorithm to solve $\boldsymbol{Q}_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b})$. Additionally, they proved that their estimates have enjoyed the Oracle properties.

Ahn et al. (2012) developed a moment-based method for selection on random effects in linear mixed models. Since their method is moment-based, the assumption of normality for the error terms is not required for their method. This attractive property results in a more robust estimation when dealing with non-normal data. Their objective function uses a second-order moment loss function with penalty terms. By optimizing their objective function, their method can effectively estimate and select random effects. They consider two types of penalty terms, including a hard thresholding operator and a sandwich-type soft thresholding penalty. Moreover, they extended their selection method to encompass the selection of fixed effects.

Pan & Shang (2018) introduced a two-stage procedure that addressed the selection of both fixed and random effects in the linear mixed model. Their method incorporates adaptive LASSO penalty terms. In the first step, the random effect is selected through the penalized restricted profile log-likelihood. In the following step, the fixed effects are determined using the profile log-likelihood function with a penalty added. They assume that $\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2\boldsymbol{V}_i(\boldsymbol{\theta}))$, where $\boldsymbol{V}_i(\boldsymbol{\theta}) = \boldsymbol{I}_{n_i} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\top$ with $\boldsymbol{\theta}$ denotes the vector consisting of $k = q(q+1)/2$ unique variance components in $\boldsymbol{D}$.

In the first stage, their proposed penalized restricted profile log-likelihood function is given as:

$$Q_R(\boldsymbol{\theta}) = p_R(\boldsymbol{\theta}) - \lambda_{1n}\sum_{j=1}^{q}\omega_{1j}|d_j|, \tag{1.6}$$

where $p_R(\boldsymbol{\theta})$ is the restricted profile log-likelihood shown in (2.7), d_j is the jth element

of $\boldsymbol{d}$, ω_{1j} is the corresponding weight of d_j, and $\lambda_{1n} \geq 0$ is the tuning parameter. Let $\boldsymbol{w}_1 = (\omega_{11}, \ldots, \omega_{1q})^\top$. They propose to use $\boldsymbol{w}_1 = 1/|\widetilde{\boldsymbol{d}}|$, where $\widetilde{\boldsymbol{d}}$ is a root-n consistent estimator of $\boldsymbol{d}$. To maximize $Q_R(\boldsymbol{\theta})$ in (1.6), they use the Newton–Raphson algorithm. A similar idea is also implemented in fixed effects selection. In the second stage, their proposed penalized profile log-likelihood is shown as:

$$Q_F(\boldsymbol{\beta}) = p_F(\boldsymbol{\beta}) - \lambda_{2n} \sum_{j=1}^{p} \omega_{2j} |\beta_j|, \tag{1.7}$$

where $p_F(\boldsymbol{\beta})$ is the profile log-likelihood defined in (2.7), $\lambda_{2n} \geq 0$ is the tuning parameter and the weight vector is suggested to define as $\boldsymbol{w}_2 = 1/|\widetilde{\boldsymbol{\beta}}|$, where $\widetilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ using the estimated covariance matrix. They also use Newton–Raphson algorithm to maximize (1.7) to find solutions. Furthermore, they showed that their proposed procedure is consistent and enjoys the oracle properties.

Li et al. (2018) developed a doubly regularized method in linear mixed models for high-dimensional longitudinal data to simultaneously select fixed and random effects. They invoked the Cholesky decomposition on the variance-covariance matrix $\boldsymbol{D}$ for random effects selection. For instance, $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is a lower triangular matrix with positive diagonal elements. Let $\boldsymbol{L}_{(k)}$ be the kth row of $\boldsymbol{L}$. If $\boldsymbol{L}_{(k)} = \boldsymbol{0}$, then the variance of the kth random effect, denoted as D_{kk}, is also zero. Therefore, instead of optimizing $\boldsymbol{D}$, they target optimizing $\boldsymbol{L}$. Their doubly regularized objective function is defined as

$$Q_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) = \ell_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) - \lambda_1 \sum_{j=1}^{p} \big|\beta_j\big| - \lambda_2 \sum_{k=2}^{q} ||\boldsymbol{L}_{(k)}||_2,$$

where $|\beta_j|$ is the absolute value of β_j, $||\boldsymbol{L}_{(k)}||_2 = \sqrt{L_{k1}^2 + \cdots + L_{kq}^2}$, and $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are tuning parameters. Moreover, when $N > p$, the restricted log-likelihood function is used in $\ell_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2)$, while the log-likelihood function for the

data is implemented in $\ell_n(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2)$ when $N < p$. Furthermore, the authors establish the large sample properties of their method for the high-dimensional setting. They introduce new regularity conditions for the diverging rates, which guarantee that the proposed method achieves both estimation and selection consistency.

1.5 Latent Variable Models

Latent variable models are a group of statistical models to investigate relationships between unobserved and observed variables. The unobserved latent variables refer to the variables that cannot be measured directly from the observed data. Even though latent variables are not observable, their information can be inferred through their relationship with observable variables. There has been a wide range of applications of latent variable models in various fields, such as psychology, economics, and social sciences, especially with applications to data analysis of longitudinal studies and repeated measures.

According to the types of observed and latent variables, the latent variable models can be classified into different models (Knott & Bartholomew, 1999). In this dissertation, we will focus on item response theory (IRT) models, a statistical model containing continuous latent variables, so-called latent traits, and categorical observed variables.

1.5.1 Common IRT Models

Psychometrics is a scientific discipline that studies testing, measurement, assessment, and related activities within psychology and education. It focuses on the development and application of psychological and educational tests and assessment tools to quantify and evaluate psychological attributes, such as personality traits, abilities,

knowledge, and so on.

In psychometrics, the item response theory (IRT) model is one of the most widely utilized latent variable models. Specifically, in the IRT model, the latent traits are continuous, while the observed variables are categorical. The IRT model is often used to analyze responses from tests and assessments. Based on the type of those responses, different IRT models can be implemented. For instance, when the test responses are binary, meaning there are only two possible outcomes, dichotomous IRT models can be utilized. In tests or assessments, the attributes of test takers that researchers are interested in, including personality traits, abilities, and knowledge, cannot be measured directly. Hence, those attributes are considered as latent traits. Through IRT models, we can deeply understand the latent traits of the test takers based on their responses to test items.

According to the number of latent traits, IRT models can be divided into unidimensional IRT models and multidimensional IRT models. Unidimensional IRT models include a single latent trait, while multidimensional IRT models consider multiple latent traits. Due to the complexity of modeling with higher dimensional latent traits, unidimensional IRT models are more frequently utilized in practice. IRT models can also be categorized based on the number of test items' responses. For multiple choice questions, the correctness of each answer is scored with either correct or incorrect. In this case, the dichotomous IRT models are used because the responses to the items are dichotomous (*i.e.*, Correct/Wrong). Polytomous IRT models are employed when there are three or more response options, such as choosing a scale from 1 to 5. In this project, we focus on unidimensional dichotomous IRT models.

We will introduce three commonly seen IRT models for binary test responses: one-parameter, two-parameter, and three-parameter logistic models. Those three

logistic models share common assumptions:

1. Unidimensinality: each test item measures only one continuous latent trait. Sometimes, it is hard to meet this assumption in practice because test takers' personalities, attitudes, and other test-taking factors can always influence test performance. However, what is required for this assumption to be met is a presence of a dominant latent factor that mainly affects test performance (Hambleton et al., 1991).

2. Local independence: given a fixed latent trait value, test takers' responses to any test item are statistically independent. It means that a test taker's response on one test item would not influence the response on another test item (Hambleton et al., 1991).

In this section, we will give notations for IRT models. For the rest of the dissertation, the notations will be consistently used.

Consider a test containing J binary test items. Assume that N subjects take this test and we only consider one latent trait $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)^\top$ for the ith subject. Let $\boldsymbol{Y}$ denote the $N \times J$ matrix of all item responses. Then, Y_{ij} represents the binary response of the ith subject on the jth test item, where 1 represents a correct answer, and 0 represents a wrong answer.

1.5.1.1 One-Parameter Logistic Model

The one-parameter logistic model, also called the Rasch model, is one of the most popular IRT models in applications. It has the simplest form of IRT models. This model has a mathematical form of the probability of ith subject answering the jth item correctly as below:

$$P(Y_{ij} = 1|\theta_i, b_j) = \frac{exp(\theta_i + b_j)}{1 + exp(\theta_i + b_j)},$$

where b_j is an item parameter that represents easiness of jth test item and θ_i indicates the latent trait of ith subject. Since the coefficient of θ_i is 1, it indicates that the Rache model treats each item with the same ability of discrimination. Therefore, only the easiness level of items and abilities of subjects affect the probability of correct answers simultaneously.

Under local independence assumption, the joint likelihood of responses matrix $\boldsymbol{Y} = (\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_N)^\top$ is

$$L(\boldsymbol{Y}, \boldsymbol{\theta}; \boldsymbol{b}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \frac{exp(\theta_i + b_j)}{1 + exp(\theta_i + b_j)}^{Y_{ij}} \frac{1}{1 + exp(\theta_i + b_j)}^{(1-Y_{ij})}.$$

Then, the marginal likelihood of $\boldsymbol{Y}$ given a latent trait and an item parameter $\boldsymbol{b}$ can be obtained by integrating $\boldsymbol{\theta}$ out:

$$L(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{b}) = \prod_{i=1}^{N} \left[\int_{\theta_i} \prod_{j=1}^{J} \frac{exp(\theta_i + b_j)}{1 + exp(\theta_i + b_j)}^{Y_{ij}} \frac{1}{1 + exp(\theta_i + b_j)}^{(1-Y_{ij})} f(\theta_i) d\theta_i \right],$$

where $f(\theta_i)$ is the probability density function (pdf) of the ith subject's latent trait.

1.5.1.2 Two-Parameter Logistic Model

The two-parameter logistic model (2PL model) contains two item parameters: discrimination and easiness. Discrimination parameter $\boldsymbol{a}$ measures the differential capability of an item. An item with a high discrimination parameter value indicates its high ability to distinguish test takers. The probability of correct responses would increase faster on items as the latent trait increases. Moreover, the easiness parameter measures the easiness of an item.

For the ith test taker on the jth item, the probability of getting a correct response Y_{ij} is modeled as

$$P(Y_{ij} = 1 | \theta_i; a_j, b_j) = \frac{exp(a_j\theta_i + b_j)}{1 + exp(a_j\theta_i + b_j)}, \tag{1.8}$$

where a_j is the discrimination parameter, b_j is the easiness parameter for $j = 1, \ldots, J$, and θ_i represents the latent trait for subject i.

Under local independence assumption, the joint likelihood of responses matrix $\boldsymbol{Y}$ is

$$L(\boldsymbol{Y}, \boldsymbol{\theta}; \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \frac{exp(a_j\theta_i + b_j)}{1 + exp(a_j\theta_i + b_j)}^{Y_{ij}} \frac{1}{1 + exp(a_j\theta_i + b_j)}^{(1-Y_{ij})}$$

and the marginal likelihood $L(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b})$ is

$$L(\boldsymbol{Y}; \boldsymbol{\theta}; \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{N} \left[\int_{\theta_i} \prod_{j=1}^{J} \frac{exp(a_j\theta_i + b_j)}{1 + exp(a_j\theta_i + b_j)}^{Y_{ij}} \frac{1}{1 + exp(a_j\theta_i + b_j)}^{(1-Y_{ij})} f(\theta_i) d\theta_i \right], \tag{1.9}$$

where $\boldsymbol{a} = (a_1, \ldots, a_J)^\top$ is a vector of discrimination parameters, $\boldsymbol{b} = (b_1, \ldots, b_J)^\top$ is a vector of easiness parameters of J test items, and $f(\theta_i)$ represents the probability density function of θ_i.

1.5.1.3 Three-Parameter Logistic Model

The three-parameter logistic model (Birnbaum, 1968) extends the two-parameter logistic model by introducing an extra item parameter: the guessing parameter, denoted by $\boldsymbol{c}$. The guessing parameter describes the probability that a test takes a correct response by guessing alone. Guessing usually happens when a test taker is unsure about the correct answer and still tries to give an answer. (Hutchinson, 1991; Maris, 1995; Martín et al., 2006). The probability of getting a correct answer from ith subject on jth test item is then given by:

$$P(Y_{ij} = 1 | \theta_i; a_j, b_j, c_j) = c_j + (1 - c_j) \frac{exp(a_j\theta_i + b_j)}{1 + exp(a_j\theta_i + b_j)},$$

where a_j is the discrimination parameter, b_j is the easiness parameter, c_j is the guessing parameter each test item, and θ_i represents the latent trait for subject i.

1.5.2 Latent Regression IRT Models

The latent regression IRT model is a type of extension to the traditional IRT model, where the inclusion of observed covariates as predictors allows for investigating their effects on the conditional distribution of the latent trait. By incorporating regression onto latent traits, latent regression IRT models provide a way to uncover hidden influential factors from the data and make more accurate predictions. In other words, the latent regression IRT model incorporates external predictors of test takers to explore how they might affect the latent traits and the observed responses to test items. Based on research interests, a variety of external predictors may be considered, such as demographic variables, including age, gender, education, and others, for test takers in large-scale assessment programs.

Latent regression IRT models assume that a set of predictors could be directly linked to the latent traits through a linear relationship. Additionally, the conditional distribution of the latent traits follows a normal distribution when the dimension of the latent trait is one or a multivariate normal distribution when the dimension of the latent traits exceeds one. Generally, each individual is characterized by a mean vector determined by the observed covariates and regression parameters.

Maximum likelihood estimation of latent regression IRT models involves calculating integrals in the likelihood function, which presents a significant challenge due to the difficulty in finding explicit solutions. Consequently, instead of finding exact solutions, approximations are needed to be utilized during estimations. Therefore, many approximation approaches have been studied in the literature. In large-scale educational assessment programs, the estimations of latent regression models are usually implemented in two steps (von Davier & Sinharay, 2013). The first step is to make an estimation on the item parameters based on a unidimensional IRT model. At this step, the predictors are ignored when fitting the model. The second step is to

assess the latent regression and variance parameters by assuming the item parameters are held constant. In the second step, there are several methods can be used. One option is to utilize an expectation-maximization (EM) algorithm with a second-order Laplace approximation method, as described by Thomas (1993). Another option is to apply stochastic approximation, as suggested by von Davier & Sinharay (2010).

In addition to such a two-step estimation approach, it is feasible to estimate the item and regression parameters simultaneously by utilizing adaptive quadrature, stochastic approximations, or a Laplace approximation (Chalmers, 2015; Harrell, 2015; Raudenbush et al., 2000). However, due to the heavy computational burden, those methods are not commonly used in large-scale assessment programs. Andersson & Xin (2021) proposed to use a second-order Laplace approximation of the likelihood to estimate latent regression IRT models. Their method can be employed on the data with categorical observed variables. Additionally, their method can estimate all parameters simultaneously. They showed their approximation approach significantly improved over the first-order Laplace approximation in terms of bias. Furthermore, their method exhibits high computational efficiency, particularly in large sample sizes and a substantial number of items.

Variable selection in latent variable models has been a new research area in psychometrics. In large-scale assessment programs, besides a large number of item responses from students, a wealth of additional information about examinees can also be collected, such as students' demographic information, academic record and school experience, affective disposition, and more. Consequently, an essential question arises: can we identify the factors significantly related to the latent trait being considered during the assessment? For example, in a mathematics test, which tends to quantify students' mathematics ability, we wonder whether we can ascertain the important factors that affect their mathematical skills. Are the number of days

absent from school correlated with their math abilities? Does computer access at home influence their mathematical ability? To answer such questions, the current operational procedures to perform such variable selection can be done in two steps. First, we fit a unidimensional IRT model to estimate item parameters and the latent variable, disregarding the influence of additional covariates. Second, we employ a variable selection method, such as stepwise selection or LASSO, on a set of factors to choose the most important factors affecting the latent trait. Even though such a two-step procedure is easy to implement in practice, it produces biased estimates of regression parameters and tends to yield unsatisfactory prediction performance. To enhance the accuracy and reliability of variable selection results, we propose an innovative method for variable selection.

Chapter 2

Random Effects Selection

2.1 New Methodology

A significant body of literature has been dedicated to selecting random effects, with a predominant focus on the selection of individual random effects. However, in the context of this dissertation, we present a novel methodology that specifically targets the selection of grouped random effects as opposed to individual random effects.

Recall from (1.5), the linear mixed model can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, \ldots, m, \tag{2.1}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})$ is a $n_i \times 1$ vector of observations for subject i, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i})^\top$ is a $n_i \times p$ design matrix of fixed-effect covariates for subject i, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of fixed-effect coefficients, $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{in_i})^\top$ is a $n_i \times q$ design matrix of random-effect covariates for subject i, $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{iq})$ is a q vector of random-effect coefficients and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{in_i})$ is a $n_i \times 1$ random error vector. Like ML and REML estimations, we also assume that $\boldsymbol{b}_i$ and $\boldsymbol{\epsilon}_i$ follow a normal distribution, respectively. To be specific, $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{D})$ where the matrix

$\boldsymbol{D}$ is a symmetric and positive semi-definite matrix, and $\boldsymbol{\epsilon}_i \overset{iid}{\sim} N(\mathbf{0}, \sigma^2 \boldsymbol{I}_{\boldsymbol{n_i}})$ where $\boldsymbol{I}_{\boldsymbol{n_i}}$ is an $n_i \times n_i$ identity matrix with all 1's on the diagonal. Therefore, it can be derived that the responses of the ith cluster $\boldsymbol{y}_i$ also follows a multivariate normal distribution with mean vector $\boldsymbol{X}_i\boldsymbol{\beta}$ and variance-covariance $\sigma^2\boldsymbol{V}_i$ with $\boldsymbol{V}_i = \boldsymbol{I}_{n_i} + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\top$. It has been discussed that the selection rule of random effects is based on the underlying structure of the variance-covariance matrix $\boldsymbol{D}$. Specifically, if the jth random effect is unimportant, then its corresponding variance is zero. Consequently, this is equivalent to setting all elements in both the jth column and the jth row of matrix $\boldsymbol{D}$ to zero.

Let $\boldsymbol{\theta}$ denote the parameter vector of the linear mixed model in (2.1) with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{D})^\top$. One common method to estimate the parameter vector $\boldsymbol{\theta}$ is by the method of maximum likelihood (Laird & Ware, 1982), which gives the ML estimators, also called MLE. The MLE can be obtained by maximizing the log-likelihood function:

$$l(\boldsymbol{\theta}, \sigma^2) = -\frac{1}{2}N\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{m}\log\boldsymbol{V}_i - \frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{2.2}$$

It is known that the log-likelihood function in (2.2) is maximized at

$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{m}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \tag{2.3}$$

By Lindstrom & Bates (1988) and Wolfinger et al. (1994), if we substitute (2.3) into the log-likelihood function in (2.2), we will have an equivalent objective function but with σ^2 eliminated, that called variance-profile log-likelihood function. To make it simple, we call it the profile log-likelihood function. Then, the profile log-likelihood function of the model in (1.5) is given by:

$$p_F(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{m}\log|\boldsymbol{V}_i| - \frac{N}{2}\sum_{i=1}^{m}((\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top\boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})). \tag{2.4}$$

If the variance-covariance matrix of the random effects $\boldsymbol{D}$ is known, the ML estimate of the coefficients of the fixed effects can be found by the generalized least squares:

$$\widehat{\boldsymbol{\beta}}^{ML} = \left(\sum_{i=1}^{m} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{y}_i \right).$$

Since there is no simple expression for the ML estimator of the covariance components, e.g., $\boldsymbol{D}$, it requires some iterative techniques, such as the EM algorithm and Newton-Raphson algorithm, to find the solutions. Moreover, since ML estimation does not account for the loss of degrees of freedom incurred during the estimation of the fixed effects, the ML estimates of the covariance component are known to be biased. Therefore, restricted ML (REML) estimation would be preferred in the estimation of covariance. The restricted log-likelihood is defined as:

$$l_R(\boldsymbol{\theta}, \sigma^2) = l(\boldsymbol{\theta}, \sigma^2) - \frac{1}{2} \log \left| \frac{1}{\sigma^2} \sum_{i=1}^{m} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right|, \tag{2.5}$$

where $l(\boldsymbol{\theta}, \sigma^2)$ is the log-likelihood function from (2.2). By maximizing (2.5) with respect to σ^2, the REML estimate of σ^2 can be obtained as:

$$\widehat{\sigma}^2_{REML} = \frac{1}{N-p} \sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})^\top \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}). \tag{2.6}$$

Then, the profile restricted log-likelihood function can be obtained by replacing σ^2 defined in (2.5) with $\widehat{\sigma}^2_{REML}$ defined in (2.6), which is given by:

$$p_R(\boldsymbol{\theta}) = -\frac{1}{2} \log \left| \sum_{i=1}^{m} \boldsymbol{X}_i{}^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right| - \frac{1}{2} \sum_{i=1}^{m} \log |\boldsymbol{V}_i| - \frac{1}{2} (N-p) \log \left[\sum_{i=1}^{m} \boldsymbol{r}_i{}^\top \boldsymbol{V}_i^{-1} \boldsymbol{r}_i \right] \tag{2.7}$$

with $\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}$.

REML estimation accounts for the degree of freedom lost by estimating the fixed

effects; thus, it provides unbiased estimates of the covariance components. Due to such attractive property, we propose to use the restricted log-likelihood function in our objective function. To make optimization iteration converge in fewer steps, we also adopt the profile restricted log-likelihood function by replacing σ^2 with $\widehat{\sigma}^2_{REML}$ defined in (2.6). Also, optimizing the profile log-likelihood needs simpler derivatives and has more consistent convergence (Lindstrom & Bates, 1988). To guarantee the positive-definiteness of the estimated variance-covariance matrix $\boldsymbol{D}$ during computation, we apply the Cholesky decomposition, *i.e.*, $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is a lower triangular matrix with non-negative diagonal entries and $\boldsymbol{L}^\top$ is its conjugate transpose. We take advantage of Cholesky decomposition because it is numerically stable and accurate (Turing, 1948). Therefore, instead of estimating $\boldsymbol{D}$, we focus on estimating the decomposed matrix $\boldsymbol{L}$. Additionally, since we are especially interested in selecting important grouped random effects, we adopt a group LASSO type penalty (Yuan & Lin, 2006) in the framework of linear mixed models.

Assume there are G groups of random effects, and the gth group has u_g covariates for $g = 1, \ldots, G$. We also call u_g as the group size for the gth group. For $g = 1, \ldots, G$, we define a vector $\boldsymbol{L}_g$ as a vectorization of all the row vectors of matrix $\boldsymbol{L}$ corresponding to the gth group:

$$\boldsymbol{L}_g = vec(L_{g*}),$$

where L_{g*} corresponds to all row vectors that belong to the gth group.

For illustration on finding $\boldsymbol{L}_g$, let us assume we obtain a 5×5 matrix $\boldsymbol{L}$ after

decomposing on the variance-covariance matrix $\boldsymbol{D}$, which is

$$\begin{pmatrix} l_{11} & 0 & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} & 0 \\ l_{51} & l_{52} & l_{53} & l_{54} & l_{55} \end{pmatrix}.$$

We also assume that the first three rows belong to the first group while the last two rows correspond to the second group. Next, we can obtain $\boldsymbol{L}_g$ for $g = 1, 2$ as below

$$\boldsymbol{L}_1 = (l_{11}, 0, 0, 0, 0, l_{21}, l_{22}, 0, 0, 0, l_{31}, l_{32}, l_{33}, 0, 0)^\top \tag{2.8}$$

and

$$\boldsymbol{L}_2 = (l_{41}, l_{42}, l_{43}, l_{44}, 0, l_{51}, l_{52}, l_{53}, l_{54}, l_{55})^\top. \tag{2.9}$$

Then, we define the penalty term as

$$\sum_{g=1}^{G} \sqrt{u_g} ||\boldsymbol{L}_g||_2, \tag{2.10}$$

where G is the number of groups of random effects, u_g is the number of variables within gth group, and $||\cdot||_2$ represents L_2 norm. The L_2 norm, also called the Euclidean norm, is the square root of the sum of the squares of the vector's components. For example, if a vector $x = (x_1, x_2)^\top$, then its L_2 norm can be calculated by $||x||_2 = \sqrt{x_1^2 + x_2^2}$. If we use $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ in (2.8) and (2.9) as an example, we can calculate

$$||\boldsymbol{L}_1||_2 = \sqrt{l_{11}^2 + l_{21}^2 + l_{22}^2 + l_{31}^2 + l_{32}^2 + l_{33}^2} \tag{2.11}$$

and

$$||\boldsymbol{L}_2||_2 = \sqrt{l_{41}^2 + l_{42}^2 + l_{43}^2 + l_{44}^2 + l_{51}^2 + l_{52}^2 + l_{53}^2 + l_{54}^2}. \tag{2.12}$$

Therefore, in this example, we can obtain the penalty term in (2.10) as

$$\sum_{g=1}^{2} \sqrt{u_g}||\boldsymbol{L}_g||_2 = \sqrt{3} \cdot ||\boldsymbol{L}_1||_2 + \sqrt{2} \cdot ||\boldsymbol{L}_2||_2,$$

where $||\boldsymbol{L}_1||_2$ and $||\boldsymbol{L}_2||_2$ are calculated as (2.11) and (2.12), respectively, and $\mu_1 = 3$ and $\mu_2 = 2$ because there are 3 and 2 covariates in each group.

Finally, we propose to minimize the following objective function:

$$Q(\boldsymbol{L}) = -p_R(\boldsymbol{L}) + \lambda \sum_{g=1}^{G} \sqrt{u_g}||\boldsymbol{L}_g||_2, \tag{2.13}$$

where $p_R(\boldsymbol{L})$ is the restricted profile log-likelihood function defined in (2.5), G is the number of groups of random effects, u_g is the number of variables within gth group, $\boldsymbol{L}_g$ is the vector of elements associated with gth group, and $\lambda \geq 0$ is a tuning parameter. By leveraging the property of the group LASSO penalty, which enables the coefficients of variables within a group to be shrunk to zero, our method extends this capability to the row vectors of matrix $\boldsymbol{L}$ corresponding to the same group, effectively shrinking it towards zero. Let $\boldsymbol{D}_g = (d_{g1}, \cdots, d_{g\mu_g})^\top$, where d_{gj} indicates the diagonal elements of the matrix $\boldsymbol{D}$ corresponding to the gth group. Then, for any given g, we have the following selection rule at the group level:

$$\boldsymbol{L}_g = \boldsymbol{0} \Longleftrightarrow \boldsymbol{D}_g = \boldsymbol{0} \Longleftrightarrow g\text{th group is not important}$$

If all row vectors of matrix $\boldsymbol{L}$ corresponding to the gth group are successfully shrunk to zero, it implies that the corresponding variances in matrix $\boldsymbol{D}$ are also shrunk to zero. This observation indicates that the considered group does not significantly

contribute as a random effect group.

2.2 Computational Algorithm

Solving $\boldsymbol{L}$ directly from (2.13) is challenging. Inspired by Lin & Zhang (2006) and Li et al. (2018), we utilize a transformation to reframe the original objective function into a more easily solvable equivalent form.

Proposition 2.1 *Given any $\boldsymbol{\beta}$ and λ, consider two objective functions as follows:*

$$Q_1(\boldsymbol{L}|\boldsymbol{\beta}) = -p_R(\boldsymbol{L}|\boldsymbol{\beta}) + \lambda \sum_{g=1}^{G} \sqrt{u_g}||\boldsymbol{L}_g||_2 \tag{2.14}$$

$$Q_2(\boldsymbol{L}, \boldsymbol{\gamma}|\boldsymbol{\beta}) = -p_R(\boldsymbol{L}|\boldsymbol{\beta}) + \sum_{g=1}^{G} {\gamma_g}^2 + \frac{\lambda^2}{4} \sum_{g=1}^{G} \frac{u_g}{\gamma_g^2}||\boldsymbol{L}_g||_2^2 \tag{2.15}$$

Let $\widehat{\boldsymbol{L}}_g$ be the minimizer of (2.14) and $(\widetilde{\boldsymbol{L}}_g, \widetilde{\gamma}_g)$ be the minimizer of (2.15). Then, it can be proved:

$$\widehat{\boldsymbol{L}}_g = \widetilde{\boldsymbol{L}}_g, \qquad g = 1, \ldots, G$$

$$\widetilde{\gamma}_g^2 = \frac{\lambda}{2}\sqrt{u_g}||\widetilde{\boldsymbol{L}}_g||_2, \qquad g = 1, \ldots, G \tag{2.16}$$

Proposition 2.1 states that instead of minimizing (2.14) with respect to $\boldsymbol{L}$ directly to find the solution for $\boldsymbol{L}$, it is equivalent to minimizing (2.15) iteratively between $\boldsymbol{L}_g$ and γ_g. When γ_g is fixed, the objective function (2.15) looks similar to a generalized ridge regression problem. Thus, we can use the Newton-Raphson algorithm to solve it. When $\boldsymbol{L}_g$'s are fixed, $\widetilde{\gamma}_g$ can be obtained based on (2.16).

The Newton-Raphson algorithm is a widely adopted iterative method to find the optimizer of a function. Assume we want to optimize a function $f(\boldsymbol{t})$ with respect to

$\boldsymbol{t}$. Denote $\boldsymbol{t}^{(r)}$ as parameter estimates at rth iteration, $\boldsymbol{t}^{(r+1)}$ as updated parameter estimates at $(r+1)$th iteration, $\boldsymbol{g}(\boldsymbol{t})^{(r)}$ as gradient vector of $f(\boldsymbol{t})$ with respect to $\boldsymbol{t}$, $\boldsymbol{H}(\boldsymbol{t})^{(r)}$ as the Hessian matrix of $f(\boldsymbol{t})$ with respect to $\boldsymbol{t}$, and $\lambda^{(r)}$ as step length. Then, the process of the Newton-Raphson algorithm is repeated as

$$\boldsymbol{t}^{(r+1)} = \boldsymbol{t}^{(r)} - \lambda^{(r)} \boldsymbol{H}(\boldsymbol{t})^{(r)^{-1}} \boldsymbol{g}(\boldsymbol{t})^{(r)}, \qquad r = 0, 1, \cdots$$

until a convergence criterion is reached. Newton-Raphson algorithm tends to converge fast, especially when initial estimates are close to the true solution. Also, it is easy to implement if gradient and Hessian matrix are available. Therefore, the Newton-Raphson algorithm is applied during the optimization procedure in our method.

We derive the gradient and Hessian matrix with respect to $\boldsymbol{L}$ for the term $p_R(\boldsymbol{L}|\boldsymbol{\beta})$ in (2.15). The following notations will be used in the derivation. Let

$$\begin{aligned}
\boldsymbol{v}_i &= \boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{r}_i \text{ be a } q \times 1 \text{ vector with } \boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}, \\
\boldsymbol{B}_i &= \boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{Z}_i \otimes \boldsymbol{v}_i \boldsymbol{v}_i^\top \text{ be a } q^2 \times q^2 \text{ matrix}, \\
\boldsymbol{C}_i &= \boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \text{ be a } q \times p \text{ matrix}, \\
\boldsymbol{H}_i &= \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X} \text{ be a } p \times p \text{ matrix}, \\
\widetilde{\boldsymbol{L}} &= \text{diag}(\boldsymbol{L}, \ldots, \boldsymbol{L}) \text{ be a } q^2 \times q^2 \text{ matrix},
\end{aligned}$$

and $\otimes$ denote the Kronecker product. By Lindstrom & Bates (1988), we can obtain

the first and second derivative of the first term in (2.7) as

$$\begin{aligned}\frac{\partial \log|\sum_{i=1}^{m}\boldsymbol{X}_i{}^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{X}_i|}{\partial \mathrm{vec}(\boldsymbol{L})} &= \widetilde{\boldsymbol{L}}\left(\frac{\partial \log|\sum_{i=1}^{m}\boldsymbol{X}_i{}^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{X}_i|}{\partial \mathrm{vec}(\boldsymbol{D}^{\top})}+\frac{\partial \log|\sum_{i=1}^{m}\boldsymbol{X}_i{}^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{X}_i|}{\partial \mathrm{vec}\,(\boldsymbol{D})}\right)\\ &= \widetilde{\boldsymbol{L}}\left[-\mathrm{vec}\left(\sum_{i}^{m}\boldsymbol{C}_i\boldsymbol{H}^{-1}\boldsymbol{C}_i^{\top}\right)-\mathrm{vec}\left(\sum_{i}^{m}\boldsymbol{C}_i\boldsymbol{H}^{-1}\boldsymbol{C}_i^{\top}\right)\right],\end{aligned} \tag{2.17}$$

and

$$\begin{aligned}\frac{\partial^2 \log|\sum_{i=1}^{m}\boldsymbol{X}_i{}^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{X}_i|}{\partial(\boldsymbol{L}^{(k)})^{\top}\partial\boldsymbol{L}^{(j)}} &= \left(\frac{\partial \log|\boldsymbol{X}^{\top}\boldsymbol{V}^{-1}\boldsymbol{X}|}{\partial \boldsymbol{D}_{jk}}+\frac{\partial \log|\boldsymbol{X}^{\top}\boldsymbol{V}^{-1}\boldsymbol{X}|}{\partial \boldsymbol{D}_{kj}}\right)\boldsymbol{I}\\ &\quad+2\boldsymbol{L}\left(\frac{\partial^2 \log|\boldsymbol{X}^{\top}\boldsymbol{V}^{-1}\boldsymbol{X}|}{\partial \boldsymbol{D}^{[k]}\partial \boldsymbol{D}^{(j)}}+\frac{\partial^2 \log|\boldsymbol{X}^{\top}\boldsymbol{V}^{-1}\boldsymbol{X}|}{\partial(\boldsymbol{D}^{(k)})^{\top}\partial(\boldsymbol{D}^{(j)})}\right)\boldsymbol{L}^{\top},\end{aligned} \tag{2.18}$$

where $\boldsymbol{D}^{[k]}$ denotes the kth row of $\boldsymbol{D}$ and $\boldsymbol{D}^{(j)}$ denotes the jth column of $\boldsymbol{D}$. To compute (2.18), we use the following information:

$$\frac{\partial \log|\boldsymbol{X}^{\top}\boldsymbol{V}^{-1}\boldsymbol{X}|}{\partial D_{jk}} = -\sum_{i}^{m} tr\left[\boldsymbol{H}^{-1}\boldsymbol{X}_i^{\top}(\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i^{(j)}\boldsymbol{Z}_i^{(k)^{\top}}\boldsymbol{V}_i)\boldsymbol{X}_i\right],$$

and

$$\begin{aligned}\frac{\partial^2 \log|\boldsymbol{X}^{\top}\boldsymbol{V}^{-1}\boldsymbol{X}|}{\partial \mathrm{vec}(\boldsymbol{D}^{\top})^{\top}\partial \mathrm{vec}(\boldsymbol{D})} &= -\left[\sum_{i}(\boldsymbol{C}_i(\boldsymbol{H}^{-\frac{1}{2}})^{\top}\otimes\boldsymbol{C}_i(\boldsymbol{H}^{-\frac{1}{2}})^{\top})\right]\times\left[\sum_{i}(\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C}_i^{\top})\otimes((\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{C}_i^{\top})\right]\\ &\quad+2\sum_{i}\left[\boldsymbol{C}_i\boldsymbol{H}^{-1}\boldsymbol{C}_i^{\top}\otimes\boldsymbol{Z}_i^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i+\boldsymbol{Z}_i^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i\otimes\boldsymbol{C}_i\boldsymbol{H}^{-1}\boldsymbol{C}_i^{\top}\right].\end{aligned}$$

For the second term in (2.7), the first and second derivatives can be derived as

$$\begin{aligned}\frac{\partial \log|\boldsymbol{V}_i|}{\partial \mathrm{vec}(\boldsymbol{L})} &= \widetilde{\boldsymbol{L}}\left(\frac{\partial \log|\boldsymbol{V}_i|}{\partial \mathrm{vec}(\boldsymbol{D}^{\top})}+\frac{\partial \log|\boldsymbol{V}_i|}{\partial \mathrm{vec}(\boldsymbol{D})}\right)\\ &= \widetilde{\boldsymbol{L}}\left(\mathrm{vec}(\boldsymbol{Z}_i^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i)+\mathrm{vec}(\boldsymbol{Z}_i^{\top}\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i)\right),\end{aligned} \tag{2.19}$$

and

$$\begin{aligned}\frac{\partial^2 \log|\boldsymbol{V}_i|}{\partial(\boldsymbol{L}^{(k)})^\top \partial \boldsymbol{L}^{(j)}} &= \left(\frac{\partial \log|\boldsymbol{V}_i|}{\partial \boldsymbol{D}_{jk}} + \frac{\partial \log|\boldsymbol{V}_i|}{\partial \boldsymbol{D}_{kj}}\right)\boldsymbol{I} \\ &\quad + 2\boldsymbol{L}\left(\frac{\partial^2 \log|\boldsymbol{V}_i|}{\partial \boldsymbol{D}^{[k]}\partial \boldsymbol{D}^{(j)}} + \frac{\partial^2 \log|\boldsymbol{V}_i|}{\partial \boldsymbol{D}^{(k)\top}\partial \boldsymbol{D}^{(j)}}\right)\boldsymbol{L}^\top.\end{aligned} \tag{2.20}$$

We use the following information to compute (2.20):

$$\frac{\partial \log|\boldsymbol{V}_i|}{\partial \boldsymbol{D}_{jk}} = tr\left(\boldsymbol{V}_i^{-1}\boldsymbol{V}_i^{-1}\boldsymbol{Z}_i^{(j)}(\boldsymbol{Z}_i^{(k)})^\top \boldsymbol{V}_i^{-1}\right),$$

and

$$\frac{\partial^2 \log|\boldsymbol{V}_i|}{\partial \mathrm{vec}(\boldsymbol{D}^\top)^\top \partial \mathrm{vec}(\boldsymbol{D})} = -\boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{Z}_i.$$

For the third term in (2.7), the first and second derivatives can be derived as

$$\begin{aligned}\frac{\partial \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \mathrm{vec}(\boldsymbol{L})} &= \widetilde{\boldsymbol{L}}\left(\frac{\partial \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \mathrm{vec}(\boldsymbol{D}^\top)} + \frac{\partial \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \mathrm{vec}(\boldsymbol{D})}\right) \\ &= \widetilde{\boldsymbol{L}}\left(-\mathrm{vec}(\boldsymbol{v}_i\boldsymbol{v}_i^\top) - \mathrm{vec}(\boldsymbol{v}_i\boldsymbol{v}_i^\top)\right)\end{aligned} \tag{2.21}$$

and

$$\begin{aligned}\frac{\partial^2 \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial(\boldsymbol{L}^{(k)})^\top \partial \boldsymbol{L}^{(j)}} &= \left(\frac{\partial \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \boldsymbol{D}_{jk}} + \frac{\partial \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \boldsymbol{D}_{kj}}\right)\boldsymbol{I} \\ &\quad + 2\boldsymbol{L}\left(\frac{\partial^2 \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \boldsymbol{D}^{[k]}\partial \boldsymbol{D}^{(j)}} + \frac{\partial^2 \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \boldsymbol{D}^{(k)\top}\partial \boldsymbol{D}^{(j)}}\right)\boldsymbol{L}^\top\end{aligned} \tag{2.22}$$

We use the following information to compute (2.22):

$$\frac{\partial^2 \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i}{\partial \boldsymbol{D}_{jk}} = -\boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{Z}_i^{(j)}(\boldsymbol{Z}_i^{(k)})^\top \boldsymbol{V}_i^{-1}\boldsymbol{r}_i,$$

and

$$\frac{\partial^2 \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{r}_i}{\partial \text{vec}(\boldsymbol{D}^\top)^\top \partial \text{vec}(\boldsymbol{D})} = \boldsymbol{B}_i + \boldsymbol{B}_i^\top.$$

Since the variance-covariance matrix $\boldsymbol{D}$ is symmetric in our model (1.5), that is $\boldsymbol{D} = \boldsymbol{D}^\top$. The derivation formula described in (2.17) to (2.22) can be much simplified in coding.

Consequently, the detailed optimization steps can be applied as below:

- **Step 1**: Initialize $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{L}_g^{(0)}$ and $\gamma_g^{(0)}$ with some feasible values. For example, we initialize $\boldsymbol{\beta}$ with the ML estimates of fixed effects and $\boldsymbol{L}_g$ for $g = 1, \ldots, G$ with Cholesky decomposition of REML estimates on variance components. Both ML and REML estimates can be obtained by *lmer* function in R. Additionally, we initialize γ_g with constant 1's for $g = 1, \ldots, G$.

- **Step 2**: For the rth iteration, update $\boldsymbol{L}_g$ by minimizing the following function:

$$-p_R(\boldsymbol{L}|\widehat{\boldsymbol{\beta}}^{(r-1)}) + \sum_{g=1}^{G} {\gamma_g^{(r-1)}}^2 + \frac{\lambda^2}{4} \sum_{g=1}^{G} \frac{u_g}{{\gamma_g^{(r-1)}}^2} ||\boldsymbol{L}_g||_2^2 \tag{2.23}$$

  where $\widehat{\boldsymbol{\beta}}^{(r-1)}$ and $\gamma_g^{(r-1)}$ are the estimates of $\boldsymbol{\beta}$ and γ_g from the $r-1$th step. To optimize (2.23), we implement the Newton-Raphson algorithm.

- **Step 3**: Update γ_g using

$${\gamma_g^{(r)}}^2 = \frac{\lambda}{2} \sqrt{u_g} ||\boldsymbol{L}_g^{(r)}||_2$$

- **Step 4**: Update $\boldsymbol{\beta}^{(r)}$ using

$$\boldsymbol{\beta}^{(r)} = (\boldsymbol{X}^\top {\boldsymbol{V}^{(r)}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top {\boldsymbol{V}^{(r)}}^{-1} \boldsymbol{y},$$

  where $\boldsymbol{V}^{(r)} = \boldsymbol{I}_{n_i} + \boldsymbol{Z}_i \boldsymbol{D}^{(r)} \boldsymbol{Z}_i^\top$ with $\boldsymbol{D}^{(r)} = \boldsymbol{L}^{(r)} {\boldsymbol{L}^{(r)}}^\top$.

- **Step 5**: When $\max_g \left\{ \left| \boldsymbol{L}_g^{(r)} - \boldsymbol{L}_g^{(r-1)} \right| \right\}$ is smaller than a threshold value, we stop the iterations and consider it as convergence. Otherwise, let $r = r + 1$ and repeat Steps 2 to 4 until the convergence criterion mentioned above is met. In the simulation, we use 0.0001 as our threshold value.

2.3 Selection of Tuning Parameter λ

The choice of tuning parameter λ in (2.14) and (2.15) plays an important role in model performance. A proper tuning parameter would result in better selection performance. Many selection criteria have been discussed in Sections 1.2.1 and 1.4.

We adopt a modified BIC as a selection criterion that defined as

$$BIC_R = -2 \times p_R(\widehat{\boldsymbol{L}}) + log(N) \times df_R, \tag{2.24}$$

where df_R is the number of non-zero diagonal elements in the estimated variance-covariance matrix $\boldsymbol{D}$ and N is the total number of observations. In the simulations, we select the tuning parameter λ that results in the minimum value of BIC_R.

2.4 Simulation Studies

In the simulation studies, we are interested in the selection performance of the proposed method under various scenarios. We use the following measurements to measure the performance of our simulation studies: the number of zero coefficients that are correctly estimated as zero (denoted as 'CZ'), the number of non-zero coefficients that are incorrectly set to zero (denoted as 'IZ'), the number of unimportant groups which are correctly estimated as unimportant groups (denoted as 'CZ^*'), the number of important groups which are incorrectly set to zero (denoted as 'IZ^*'), the number

of non-zero coefficients that are correctly estimated as non-zero (denoted as 'CN'), the number of important groups that are correctly selected(denoted as 'CN^*'), the number of models that select non-zero individual coefficients correctly(denoted as 'F'), the number of models that select important groups correctly(denoted as 'F^*'), the frequency of selecting the correct model (denoted as 'C'), the frequency of over-selecting variables (denoted as 'O'), and the frequency of under-selecting variables (denoted as 'U').

2.4.1 Setting

We perform 100 runs for each scenario, and the median performance is recorded. We use BIC defined in (2.24) as our selection criterion to choose the best tuning parameter. We make the assumption that neither the fixed intercept nor the random intercept is included in the model. We consider three different scenarios.

Example 1: We consider $n = 50$ subjects and $n_i = 5$ observations per subject. The true model with $p = 6$ for fixed effects and $q = 7$ for random effects. The true fixed effects vector $\beta = (2, 2, 2, 0, 0, 0)^T$ and the true covariance matrix

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0.7 & 0.49 & 0 & 0 & 0 & 0 \\ 0.7 & 1 & 0.7 & 0 & 0 & 0 & 0 \\ 0.49 & 0.7 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.7 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

The first three columns correspond to one group, while the middle two columns

correspond to the second group. The last two columns belong to an unimportant group.

The response Y_{ij} follows:

$$\begin{aligned} Y_{ij} &= 2X_{ij,1} + 2X_{ij,2} + 2X_{ij,3} + 0X_{ij,4} + 0X_{ij,5} + 0X_{ij,6} \\ &\quad + b_{i1}Z_{ij,1} + b_{i2}Z_{ij,2} + b_{i3}Z_{ij,3} + b_{i4}Z_{ij,4} + b_{i5}Z_{ij,5} + \epsilon_{ij}, \end{aligned}$$

for $i = 1, \ldots, 50$ and $j = 1, \ldots, 5$.

We generate simulated data using the following assumed distributions:

- $\boldsymbol{X}_i \sim MVN(\boldsymbol{0},\ \boldsymbol{I}_p)$
- $\boldsymbol{Z}_i \sim MVN(\boldsymbol{0},\ \boldsymbol{I}_q)$
- $\boldsymbol{Y}_i \sim MVN(\boldsymbol{X}_i\boldsymbol{\beta},\ \sigma^2(\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i^\top + \boldsymbol{I}_{n_i}))$ with $\sigma^2 = 1$

Then, we simulate data with different numbers of observations for each subject, such as $n = 50$, $n = 100$, and $n = 200$.

Example 2: The setup is the same as example 1, except the true variance-covariance matrix $\boldsymbol{D}$ becomes

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0.5 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We also consider different sample sizes with $n = 50, 100,$ and 200.

Example 3: We decrease the dimension of random effects to 6. There are still three groups of random effects. The last column of $\boldsymbol{D}$ refers to an unimportant group of random effects. And the true variance-covariance $\boldsymbol{D}$ is given by

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0.1 & 0.3 & 0 & 0 & 0 \\ 0.1 & 0.2 & 0.1 & 0 & 0 & 0 \\ 0.3 & 0.1 & 0.6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Other simulation setting is the same as Example 1.

2.4.2 Results

We present a summary of the random effects selection results for each example in Table 2.1 through Table 2.6. We also compare the performance with methods from Bondell et al. (2010) and Pan & Shang (2018). For each measurement, we report the median in the tables. The 'Oracle' row in each table represents the true values of measurements from simulated datasets.

Based on the measurement results presented in Tables 2.1 and 2.2, several findings can be made. In Example 1, with subjects of 50, our method and Bondell's method exhibit commendable performance in selecting variables such as CZ, IZ, CZ^*, and IZ^*. However, Pan's method does not perform well on CZ and CZ^*, as it tends to overselect random effects. Thus, their method leads to difficulties in identifying the correct ones in this case. As the sample size increases to 100, all methods demonstrate

Table 2.1: Random effects selection results for Example 1 with seven measurements: CZ, CZ^*, IZ, IZ^*, C, U and O

Example 1	Method	CZ	IZ	CZ^*	IZ^*	C	U	O
n=50	Proposed	2	0	1	0	88	0	12
	Bondell	2	0	1	0	60	16	24
	Pan	0	0	0	0	19	8	72
n=100	Proposed	2	0	1	0	89	0	11
	Bondell	2	0	1	0	82	0	18
	Pan	2	0	1	0	90	0	10
n=200	Proposed	2	0	1	0	98	0	2
	Bondell	2	0	1	0	64	0	37
	Pan	2	0	1	0	100	0	0
	Oracle	2	0	1	0	100	0	0

Table 2.2: Random effects selection results for Example 1 with four measurements: CN, CN^*, F and F^*

Example 1	Method	CN	CN^*	F	F^*
n=50	Proposed	5	2	83	100
	Bondell	5	2	74	95
	Pan	5	2	100	100
n=100	Proposed	5	2	100	100
	Bondell	5	2	92	98
	Pan	5	2	100	100
n=200	Proposed	5	2	98	100
	Bondell	5	2	100	100
	Pan	5	2	100	100
	Oracle	5	2	100	100

Table 2.3: Random effects selection results for Example 2 with seven measurements: CZ, CZ^*, IZ, IZ^*, C, U and O

Example 2	Method	CZ	IZ	CZ^*	IZ^*	C	U	O
n=50	Proposed	2	0	1	0	83	3	14
	Bondell	2	0	1	0	54	18	28
	Pan	0	0	0	0	19	7	74
n=100	Proposed	2	0	1	0	90	2	8
	Bondell	2	0	1	0	80	0	20
	Pan	2	0	1	0	92	0	8
n=200	Proposed	2	0	1	0	98	0	2
	Bondell	2	0	1	0	52	0	48
	Pan	2	0	1	0	100	0	0
	Oracle	2	0	1	0	100	0	0

improved selection performance, with a notable increase in the frequency of selecting the correct model. Upon reaching 200 subjects, our approach and Pan's method exhibit similar and satisfactory selection performance, while Bondell's method faces convergence challenges, resulting in worse performance. It is worth noting that Table 2.2, which measures the selection of non-zero random effects, shows seemingly perfect results for all three methods. However, it is essential to consider the insights from Table 2.1 alongside Table 2.2 to gain a comprehensive understanding. For instance, when $n = 200$, Bondell's method achieves an oracle value of 100 for both F and F^*. However, this outcome results from Bondell's method preferring to overselect variables. Therefore, we need to simultaneously read Table 2.1 with Table 2.2 to have accurate insights.

From Tables 2.3 and 2.4, which present the selection results from Example 2, we observe similar performance patterns to those seen in Example 1. When $n = 50$, our method demonstrates the best performance, while Bondell's method performs

Table 2.4: Random effects selection results for Example 2 with four measurements: CN, CN^*, F and F^*

Example 2	Method	CN	CN^*	F	F^*
n=50	Proposed	5	2	97	100
	Bondell	5	2	82	98
	Pan	5	2	93	96
n=100	Proposed	5	2	98	100
	Bondell	5	2	100	100
	Pan	5	2	100	100
n=200	Proposed	5	2	100	100
	Bondell	5	2	100	100
	Pan	5	2	100	100
	Oracle	5	2	100	100

adequately, and Pan's method exhibits the weakest performance in terms of CZ, CZ^*, and C. As the sample size increases to 100, all methods show improved performance compared to the sample size of 50. When the sample size increases to 200, our method and Pan's method achieve highly accurate selection.

Example 3 involves a smaller dimension for the variance-covariance matrix and thus results in faster computational speed compared to the previous examples. Additionally, some variance components are considerably small in this case. Table 2.5 and 2.6 provide insights into the performance of each method for this example. For small sample sizes $n = 50$, our method consistently demonstrates the best performance, while Bondell's and Pan's methods tend to underselect variables. Specifically, in terms of IZ, Pan's method exhibits a large value, which indicates a significant loss of information during the selection process. For $n = 100$, all methods improve their selection performance in terms of C and U, with fewer underselected models. As the sample size increases to 200, all methods exhibit further improvements in selection

Table 2.5: Random effects selection results for Example 3 with seven measurements: CZ, CZ^*, IZ, IZ^*, C, U and O

Example 3	Method	CZ	IZ	CZ^*	IZ^*	C	U	O
n=50	Proposed	1	0	1	0	79	19	2
	Bondell	1	1	1	0	48	46	6
	Pan	1	4	1	1	25	55	20
n=100	Proposed	1	0	1	0	91	9	0
	Bondell	1	0	1	0	70	19	11
	Pan	1	0	1	0	55	45	0
n=200	Proposed	1	0	1	0	92	8	0
	Bondell	1	0	1	0	72	13	15
	Pan	1	0	1	0	80	20	0
	Oracle	1	0	1	0	100	0	0

Table 2.6: Random effects selection results for Example 3 with four measurements: CN, CN^*, F and F^*

Example 3	Method	CN	CN^*	F	F^*
n=50	Proposed	5	2	81	90
	Bondell	4	2	36	96
	Pan	4	2	25	100
n=100	Proposed	5	2	91	95
	Bondell	5	2	63	93
	Pan	5	2	55	100
n=200	Proposed	5	2	92	100
	Bondell	5	2	87	100
	Pan	5	2	80	100
	Oracle	5	2	100	100

performance across all measurements. Notably, our method outperforms the other methods in this scenario.

In conclusion, our method demonstrates superior performance when the number of subjects is small. However, as the number of subjects increases, our and Pan's methods exhibit comparable performance.

2.5 Real Data Example

We apply the proposed method to the early childhood longitudinal study for the Kindergarten Class of 1998-99 (ECLS-K), which tracks children's early school experiences from kindergarten through middle school. We enter several candidate random effects into the proposed method to choose the most important ones and use the proposed BIC_R defined in (2.24) to tune the parameter.

2.5.1 Description

ECLS-K is a longitudinal study that follows the developmental progress of the same group of children from kindergarten through 8th grade. The information is collected at seven different time points: in the fall and the spring of Kindergarten (1998-99), the fall and spring of 1st grade (1999-2000), the spring of 3rd grade (2002), the spring of 5th grade (2004), and the spring of 8th grade (2007) (*Early Childhood Longitudinal Studies Program - Kindergarten Class of 1998-99*, 1998-2007). Much information is collected, including information about the child, family, school, and their testing scores. This study collects a wide range of information, including details about the child, their family, school, and their test scores.

We focus on looking at the reading score of children ($\boldsymbol{Y}$) and how it relates to several factors of children and their families. We use $N = 1050$ subjects after data

Table 2.7: ECLS-K data set variable descriptions

Variable	Type	Description
READING	Continuous	Reading score (from $50 - 116$)
BMI	Continuous	Body mass index
LANGST	Categorical	If the child speaks English at home (1.No; 2.Yes)
PORVTY	Categorical	If the child's family is below the poverty threshold (1=No; 2=Yes)
FAMTYPE	Categorical	Family type (1=Two parents; 2=One parent; 3=other)
SIBLS	Categorical	If the child has siblings (1=No sibling; 2=One sibling; 3=More than one)

cleaning. Additionally, we choose $p = 5$ variables as fixed effects, including BMI, LANGST, PORVTY, FAMTYPE, and SIBLS, and $q = 5$ same variables as random effects. Table 2.7 provides detailed variable information, and the first category for each variable is referred to as the reference group. Consequently, we build the model for describing the reading score of the ith student at the jth measurement as

$$\begin{aligned} Y_{ij} =& \beta_1 \text{BIM}_{ij} + \beta_2 I(\text{LANGST}_{ij} = 2) + \beta_3 I(\text{PORVTY}_{ij} = 2) \\ &+ \beta_4 I(\text{FAMTYPE}_{ij} = 2) + \beta_5 I(\text{FAMTYPE}_{ij} = 3) \\ &+ \beta_6 I(\text{SIBLS}_{ij} = 2) + \beta_7 I(\text{SIBLS}_{ij} = 3) \\ &+ b_1 \text{BIM}_{ij} + b_2 I(\text{LANGST}_{ij} = 2) + b_3 I(\text{PORVTY}_{ij} = 2) \\ &+ b_4 I(\text{FAMTYPE}_{ij} = 2) + b_5 I(\text{FAMTYPE}_{ij} = 3) \\ &+ b_6 I(\text{SIBLS}_{ij} = 2) + b_7 I(\text{SIBLS}_{ij} = 3), \end{aligned}$$

where β's are the coefficients for fixed effects and b's are the coefficients for random effects.

Table 2.8: Selection and estimation results on ECLS-K data set

Random effect Variable	REML	Our method
BMI	0.18	0
LANGST (YES)	0.59	0.52
POVRTY (NO)	0.21	0.18
FAMTYPE (1 parent)	0.05	0.0003
FAMTYPE (other)	0.32	0.28
SIBLS (1 sibling)	0.12	0
SIBLS (more siblings)	0.03	0

2.5.2 Results

Our method selects three out of five random effects, which are LANGST, PORVRT, and FAMTYPE. It indicates that whether students speak English at home, whether their family is below the poverty line, and their family type are significant random effects for their reading scores, while students' BMI and whether the student has siblings are not important. Table 2.8 presents the selection results and estimation of the variance of each variable.

Chapter 3

Variable Selection in Latent Variable Models

3.1 New Methodology

Our research question is how to select important factors related to the latent variable. The two-step approach mentioned in Section 1.5.2 is an ad-hoc approach. The first step is estimating the latent traits by ignoring the covariates, and the second is applying variable selection strategies. One of the most frequently used selection methods is stepwise selection because it is easy to implement in practice. However, there are some drawbacks to using stepwise selection. First, since stepwise selection choose variables based on parameter inference (Chatfield, 1995), it may lead to biased parameter estimation. Second, stepwise selection results are not stable and reliable. The results could be affected by selection criterion, the order of input covariates, or the number of candidates parameters (Derksen & Keselman, 1992). Motivated by such problems, we propose a new method for selecting important factors. This involves incorporating a penalty term into the marginal likelihood of item responses.

In this dissertation, we only consider the unidimensional latent regression IRT

models. Let us use $\boldsymbol{Y}$ to represent the $N \times J$ matrix of item responses and $\boldsymbol{X}$ to represent the $N \times p$ matrix of covariates of N subjects. We can also use $\boldsymbol{y}_i$ to denote the item response vector for the ith subject, which has dimensions of $J \times 1$. Similarly, the covariate vector for the ith subject can be represented by $\boldsymbol{x}_i$ and has dimensions of $p \times 1$. Denote a vector of all item parameters by $\boldsymbol{\Gamma}$. Let θ_i be the latent trait for the ith subject and $P(\boldsymbol{y}_i|\theta_i;\boldsymbol{\Gamma})$ indicate the likelihood of the response vector $\boldsymbol{y}_i$ given θ_i. Due to the assumption of local independence, as we mentioned in Section 1.5.2, $P(\boldsymbol{y}_i|\theta_i;\boldsymbol{\Gamma})$ is the product of $P(y_{ij}|\theta_i;\boldsymbol{\Gamma}_j)$ for $j = 1, \ldots, J$. Equivalently,

$$P(\boldsymbol{y}_i|\theta_i;\boldsymbol{\Gamma}) = \prod_{j=1}^{J} P(y_{ij}|\theta_i;\boldsymbol{\Gamma}_j),$$

where y_{ij} is the item response on the jth test item for subject i and $\boldsymbol{\Gamma}_j$ is a vector of item parameters for the jth item.

For each subject i, it is assumed that

$$\theta_i|\boldsymbol{x_i} = \boldsymbol{x_i}^\top \boldsymbol{\beta} + e_i, \tag{3.1}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of latent regression coefficients and error term $e_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, N$. Thus, the distribution of the latent trait conditioned on fixed covariates follows a normal distribution with mean $\boldsymbol{x_i}^\top \boldsymbol{\beta}$ and variance σ^2. In other words, the mean of each subject's latent trait is defined by both individual covariates and regression coefficients. Let $\phi(\theta_i; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2)$ denote the normal density with mean $\boldsymbol{x_i}^\top \boldsymbol{\beta}$ and variance σ^2. It has the mathematics form defined as

$$\phi(\theta_i; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(\theta_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2}{2\sigma^2}\right).$$

The joint likelihood function of item responses $\boldsymbol{y}_i$ and θ_i for each subject is shown

as below:

$$L_i(\boldsymbol{y}_i, \theta_i; \boldsymbol{\Gamma}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(\boldsymbol{y_i}|\theta_i; \boldsymbol{\Gamma}) \phi(\theta_i; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2). \tag{3.2}$$

The joint likelihood function can be used to estimate latent traits and item parameters. However, it results in asymptotically inconsistent estimates for many IRT models (Y. Chen et al., 2018). Due to this limitation, the marginal likelihood function is often preferred as an alternative. In addition, since the parameter θ_i in (3.2) is not observable and with the assumption that it comes from a fixed mean vector related to $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$, the marginal likelihood function is a more appropriate choice for estimation in the proposed method.

Therefore, by integrating out the latent trait in (3.2), the marginal likelihood of a response vector $\boldsymbol{y_i}$ can be derived from (3.2), which is given by

$$L_i(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_i, \boldsymbol{x}_i) = \int P(\boldsymbol{y_i}|\theta_i; \boldsymbol{\Gamma}) \phi(\theta_i; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2) d\theta_i \tag{3.3}$$

and the marginal likelihood for all N subjects is given by

$$L(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \int \prod_{i=1}^{N} P(\boldsymbol{y_i}|\theta_i; \boldsymbol{\Gamma}) \phi(\theta_i; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2) d\theta_i. \tag{3.4}$$

By taking log on (3.3) and (3.4), the log marginal likelihood for the ith individual is obtained as

$$\log L_i(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_i, \boldsymbol{x}_i) = \log \int P(\boldsymbol{y_i}|\theta_i; \boldsymbol{\Gamma}) \phi(\theta_i; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2) d\theta_i \tag{3.5}$$

and the log marginal likelihood for all N subjects can be represented as

$$\log L(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_i, \boldsymbol{x}_i) = \sum_{i=1}^{N} \log \int P(\boldsymbol{y_i}|\theta_i; \boldsymbol{\Gamma}) \phi(\theta_i; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2) d\theta_i. \tag{3.6}$$

Evaluating the integral part in (3.6) is a big challenge in computation because it lacks tractable analytical solutions. To overcome this challenge, we apply a second-order approximation approach (Shun, 1997; Andersson & Xin, 2021), which is able to approximate the integral more accurately and efficiently under the latent regression IRT model. Specifically, the second-order Laplace approximation utilizes up to 4th order of the function to make an approximation.

Now we adapt the second-order Laplace approximation method (Andersson & Xin, 2021) to calculate the integration in (3.6) in detail. The Laplace approximation was introduced as a method to approximate the integral of the form:

$$I(J) = \int e^{-Jh(\boldsymbol{x})} d\boldsymbol{x}, \tag{3.7}$$

where $0 < J < \infty$ and $h(\boldsymbol{x})$ being a smooth function with a unique minimum at $\boldsymbol{x}_0 \in \mathbb{R}^p$.

Let $\boldsymbol{H}_0$ be a $p \times p$ matrix, $\boldsymbol{H}_0 = \left.\dfrac{\partial^2 h(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\top}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$. Then, the integral in (3.7) can be written as in Shun (1997)

$$I(J) = \left(\frac{2\pi}{J}\right)^{p/2} |\boldsymbol{H}_0|^{-1/2} e^{-Jh(\boldsymbol{x}_0)} (1 + R_J + \cdots),$$

with

$$R_J = -\frac{1}{2J}\left[\frac{1}{4}\sum_{ijkl} h^{ijkl} b_{ik} b_{jl} - \sum_{ijkrst} h^{ijk} h^{rst} \times \left(\frac{1}{4} b_{ir} b_{jk} b_{st} + \frac{1}{6} b_{ir} b_{js} b_{kt}\right)\right],$$

where $\{b_{jk}\}$ refer to the entries in $\boldsymbol{H}_0^{-1}$,

$$h^{ijk} = \frac{\partial^3 h}{\partial x_i \partial x_j \partial x_k}$$

and

$$h^{ijkl} = \frac{\partial^4 h}{\partial x_i \partial x_j \partial x_k \partial x_l}.$$

For a subject i, a multidimensional latent regression IRT model is defined by

$$\boldsymbol{\theta}_i | \boldsymbol{x}_i \sim N(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\theta}_i$ is a vector of multidimensional latent traits, $\boldsymbol{\Sigma}$ is the variance-covariance matrix of latent traits, and $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are the same as defined in (3.1). Let

$$h_i(\boldsymbol{\theta}) = -\log\left[P(\boldsymbol{y_i}|\boldsymbol{\theta};\boldsymbol{\Gamma})\varphi(\boldsymbol{\theta};\boldsymbol{x}_i^\top\boldsymbol{\beta},\boldsymbol{\Sigma})\right], \tag{3.8}$$

where $P(\boldsymbol{y_i}|\boldsymbol{\theta};\boldsymbol{\Gamma})$ represents the likelihood of response vector $\boldsymbol{y}_i$ given $\boldsymbol{\theta}$ and $\varphi(\boldsymbol{\theta};\boldsymbol{x}_i^\top\boldsymbol{\beta},\boldsymbol{\Sigma})$ represents the probability density function of $\boldsymbol{\theta}$, and let $\widehat{\boldsymbol{\theta}}_i$ be the minimizer of $h_i(\boldsymbol{\theta})$, $i = 1, \ldots, N$. Then, the proposed second-order Laplace approximation (Andersson & Xin, 2021) of the marginal likelihood of responses for the ith subject can be represented as

$$L_i^{\text{Lap2}}(\boldsymbol{\Gamma},\boldsymbol{\beta},\boldsymbol{\Sigma}|\boldsymbol{y}_i,\boldsymbol{x}_i) = (2\pi)^{p/2}\left|\frac{\partial^2 h_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\bigg|^{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_i}\right|^{-1/2} e^{-h_i(\widehat{\boldsymbol{\theta}}_i)}(1+\epsilon_0), \tag{3.9}$$

with

$$\epsilon_0 = -\frac{1}{2}\left[\frac{1}{4}\sum_{j=1}^{p}\frac{\partial^4\widehat{h}_i}{\partial\theta_j^4}b_{i,jj}^2 - \frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{p}\frac{\partial^3\widehat{h}_i}{\partial\theta_j^3}\frac{\partial^3\widehat{h}_i}{\partial\theta_k^3}(b_{i,jk}b_{i,jj}b_{i,kk} + b_{i,jk}^3)\right],$$

where $b_{i,jj}$ is the jth column and jth row entry of the inverse of a $p \times p$ matrix $\boldsymbol{H}_i = \dfrac{\partial^2 h_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\bigg|^{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_i}$, and $\widehat{h}_i = h_i(\widehat{\boldsymbol{\theta}}_i)$. By taking log of (3.9), it is straightforward to

obtain the approximated log marginal likelihood for the i subject:

$$l_i^{\text{Lap2}}(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{y}_i, \boldsymbol{x}_i) = \frac{p}{2}\log(2\pi) - \frac{1}{2}\log\left|\frac{\partial^2 h_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|^{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_i}\Bigg| - h_i(\widehat{\boldsymbol{\theta}}_i) + \log(1+\epsilon_0).$$

Given our focus on a unidimensional latent regression IRT model defined in (3.1), where only one latent variable is taken into account, the application of the second-order Laplace approximation approach becomes more straightforward. We define a function of the latent trait $h_i(\theta)$, where

$$h_i(\theta) = -\log\left[P(\boldsymbol{y_i}|\theta;\boldsymbol{\Gamma})\phi(\theta;\boldsymbol{x_i}^\top\boldsymbol{\beta},\sigma^2)\right], \tag{3.10}$$

and thus (3.5) can be written in terms of (3.10) as

$$L_i(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}_i, \boldsymbol{x}_i) = \int_\theta e^{-h_i(\theta)}d(\theta).$$

Let $\widehat{\theta}_i$ be the minimizer of $h_i(\theta)$ defined in (3.10) for $i = 1, \ldots, N$. For notation simplicity, we let $\widehat{h} = h_i(\widehat{\theta}_i)$, which represents the value of $h_i(\theta)$ when $\theta = \widehat{\theta}_i$. Additionally, we use $h_m(\widehat{\theta}_i)$ for $m = 1, \ldots, 4$ to represent the mth order of $h_i(\theta)$ respectively, when $\theta = \widehat{\theta}_i$. Consequently, by applying the second-order Laplace approximation on (3.5), the approximated marginal likelihood of responses vector $\boldsymbol{y}_i$ for each ith individual can be derived as

$$L_i^A(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y}_i, \boldsymbol{x}_i) = \sqrt{2\pi}\ \widehat{h}_2^{-\frac{1}{2}}\ e^{-\widehat{h}}\ \left(1 - \frac{\widehat{h}_4}{8\widehat{h}_2^2} + \frac{5\widehat{h}_3^2}{24\widehat{h}_2^3}\right). \tag{3.11}$$

We can derive one more step to obtain the log marginal likelihood from (3.11), which

is given by

$$\log L_i^A(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_i, \boldsymbol{x}_i) = \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\widehat{h}_2) - \widehat{h} + \log\left(1 - \frac{\widehat{h}_4}{8\widehat{h}_2^2} + \frac{5\widehat{h}_3^2}{24\widehat{h}_2^3}\right). \quad (3.12)$$

Therefore, the log marginal likelihood of responses for all N individuals is defined as

$$\log L^A(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{N} \log L_i^A(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_i, \boldsymbol{x}_i).$$

In addition to incorporating the likelihood function, our method involves adding a shrinkage penalty on latent regression coefficients $\boldsymbol{\beta}$. Specifically, we choose the group LASSO type penalty, which takes L_2 norms on coefficients of grouped variables. This choice is motivated by the predominant presence of factors, which are categorical variables, in large-scale educational assessment programs. As mentioned in Section 1.3.7, group LASSO shrinks variables at the group level instead of the individual level. The group LASSO penalty holds a distinct advantage in effectively selecting categorical variables. By utilizing this advantageous property, our approach adopted the group LASSO penalty to improve the identification of factors that are significantly associated with the latent trait.

Assume there are m groups of predictors and the number of coefficients in the lth group is p_l with $l = 1, \ldots, m$. Finally, we propose to minimize the following objective function

$$Q(\boldsymbol{\beta}) = -\sum_{i=1}^{N} \log L_i^A(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_i, \boldsymbol{x}_i) + \lambda \sum_{l=1}^{m} \sqrt{p_l} ||\boldsymbol{\beta}^{(l)}||_2, \quad (3.13)$$

where $\log L_i^A(\boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}_i, \boldsymbol{x}_i)$ is the approximated marginal likelihood defined in (3.12), $\boldsymbol{\beta}^{(l)}$ represents latent regression coefficients associated with the lth group, and $\lambda \geq 0$ is the tuning parameter. The next step is to minimize (3.13) with respect to $\boldsymbol{\beta}$ to find the best latent regression coefficients.

3.2 Explicit Formulation for 2PL IRT Models

We present the explicit formulation for a two-parameter logistic (2PL) model as described in Section 1.5.1.2. The 2PL model is particularly useful for analyzing binary item responses, such as those encountered in multiple-choice questions, where the response options are either correct or incorrect. In general, the 2PL model allows items to vary in terms of their easiness and their ability to discriminate. It also assumes a zero probability of a correct guess on an item. Therefore, two parameters in the model refer to the easiness parameter ($\boldsymbol{b}$) and the discrimination parameter ($\boldsymbol{a}$). In some literature, the easiness parameter is often represented as the difficulty parameter by changing its sign, and the two ways of the formulation are equivalent.

We consider a 2PL model with a latent regression structure for the following derivation. Assume there are J test items and N test takers. The probability that the ith test taker whose latent trait is described by θ_i will answer the test item j correctly has been given in (1.8). Recall (1.9) and (3.1), the marginal likelihood of test responses $\boldsymbol{Y}$ for N test takers is

$$
\begin{aligned}
&L(\boldsymbol{Y}|\boldsymbol{\theta};\boldsymbol{a},\boldsymbol{b}) \qquad (3.14)\\
&= \prod_{i=1}^{N}\left[\int_{\theta_i}\prod_{j=1}^{J}\frac{exp(a_j\theta_i+b_j)}{1+exp(a_j\theta_i+b_j)}^{Y_{ij}}\frac{1}{1+exp(a_j\theta_i+b_j)}^{(1-Y_{ij})}\phi(\theta_i;\boldsymbol{x_i}^\top\boldsymbol{\beta},\sigma^2)d\theta_i\right],
\end{aligned}
$$

where $\boldsymbol{a}=(a_1,\ldots,a_J)^\top$ is a vector of discrimination parameters, $\boldsymbol{b}=(b_1,\ldots,b_J)^\top$ is a vector of easiness parameters of J test items, and $\phi(\theta_i;\boldsymbol{x_i}^\top\boldsymbol{\beta},\sigma^2)$ represents the probability density function of θ_i. Based on the assumption (3.1), the probability density function of a normal distribution with mean zero and variance σ^2 is used for $\phi(\theta_i;\boldsymbol{x_i}^\top\boldsymbol{\beta},\sigma^2)$, that is given by

$$
\phi(\theta_i;\boldsymbol{x_i}^\top\boldsymbol{\beta},\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}exp\left\{-\frac{(\theta_i-\boldsymbol{x_i}^\top\boldsymbol{\beta})^2}{2\sigma^2}\right\}.
$$

To adopt the second-order Laplace approximation on the integral in (3.14), we first write down the $h_i^*(\theta)$ based on (3.10) in the following form

$$\begin{aligned} h_i^*(\theta) &= -\log\left[\prod_{j=1}^{J}\frac{\mathrm{e}^{a_j\theta_i+b_j}}{1+\mathrm{e}^{a_j\theta_i+b_j}}^{Y_{ij}}\frac{1}{1+\mathrm{e}^{a_j\theta_i+b_j}}^{1-Y_{ij}}\phi(\theta_i;\boldsymbol{x_i}^\top\boldsymbol{\beta},\sigma^2)\right] \\ &= -\log\left\{\prod_{j=1}^{J}\frac{\mathrm{e}^{a_j\theta_i+b_j}}{1+\mathrm{e}^{a_j\theta_i+b_j}}^{Y_{ij}}\frac{1}{1+\mathrm{e}^{a_j\theta_i+b_j}}^{1-Y_{ij}}\frac{1}{\sigma\sqrt{2\pi}}exp\left[-\frac{(\theta_i-\boldsymbol{x}_i^\top\boldsymbol{\beta})^2}{2\sigma^2}\right]\right\} \\ &= -\sum_{j=1}^{J}\left[Y_{ij}\log\frac{\mathrm{e}^{a_j\theta_i+b_j}}{1+\mathrm{e}^{a_j\theta_i+b_j}}+(1-Y_{ij})\log\frac{1}{1+\mathrm{e}^{a_j\theta_i+b_j}}\right] \\ &\quad -\left[-\frac{1}{2}\log(2\pi)-\frac{1}{2}\log\sigma^2-\frac{1}{2\sigma^2}(\theta_i-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right]. \end{aligned} \tag{3.15}$$

Hence, we can derive the first derivative and even higher orders of $h_i^*(\theta)$ with respect to θ. The first derivative of $h_i^*(\theta)$ are derived as

$$\frac{\partial h_i^*(\theta)}{\partial\theta}=-\sum_{j=1}^{J}\frac{a_j\cdot\left[\left(\mathrm{e}^{b_j}Y_{ij}-\mathrm{e}^{b_j}\right)\mathrm{e}^{a_j\theta_i+Y_{ij}}\right]}{\mathrm{e}^{a_j\theta_i+b_j}+1}+\frac{\theta_i-\boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma^2}. \tag{3.16}$$

Then, we obtain the higher order, including 2nd, 3rd and 4th order, of $h_i^*(\theta)$ with respect to θ via continuing to take derivative on (3.16), which are presented as

$$\frac{\partial^2 h_i^*(\theta)}{\partial^2\theta} = \sum_{j=1}^{J}\frac{a_j^2\mathrm{e}^{a_j\theta_i+b_j}}{\left[\mathrm{e}^{a_j\theta_i+b_j}+1\right]^2}+\frac{1}{\sigma^2} \tag{3.17}$$

$$\frac{\partial^3 h_i^*(\theta)}{\partial^3\theta} = -\sum_{j=1}^{J}\frac{a_j^3\cdot\left[\mathrm{e}^{a_j\theta_i+b_j}-1\right]\mathrm{e}^{a_j\theta_i+b_j}}{\left[\mathrm{e}^{a_j\theta_i+b_j}+1\right]^3} \tag{3.18}$$

$$\frac{\partial^4 h_i^*(\theta)}{\partial^4\theta} = \sum_{j=1}^{J}\frac{a_j^4\mathrm{e}^{a_j\theta_i+b_j}\cdot\left[\mathrm{e}^{2a_j\theta_i+2b_j}-4\mathrm{e}^{a_j\theta_i+b_j}+1\right]}{\left(\mathrm{e}^{a_j\theta_i+b_j}+1\right)^4}. \tag{3.19}$$

Since $\widehat{\theta}_i$ is the minimizer of $h_i{}^*(\theta)$ defined in (3.15) and it is needed in approximating

the marginal likelihood function, we employ an optimization method on ${h_i}^*(\theta)$ to find $\widehat{\theta}_i$ for $i = 1, \ldots, N$. The detail will be described in (3.3). Once $\widehat{\theta}_i$ is obtained for each subject, we are able to plug in its value into the first to the fourth order of $h_i^*(\theta)$ derived in (3.16) to (3.19) and then approximate the marginal likelihood based on (3.12) for N subjects on J test items for a 2PL model.

To estimate other parameters, including item parameters $\boldsymbol{a}$, $\boldsymbol{b}$, and σ, the gradients of the objective function with respect to those parameters are required for computation. Since the explicit forms of Hessian matrices are considerably challenging to find in our case, we employ a modified Newton-Raphson algorithm by approximating the Hessian matrix based on information from gradients. Therefore, gradients play an important role in the optimization procedure. The general optimization steps will be described in Section 3.3. In the following part of this section, we provide explicit forms of gradients with respect to different parameters, which are utilized in computing the vectors of the gradient.

Let

$$\widehat{h} = h_i^*(\theta), \quad \widehat{h}_2 = \frac{\partial^2 h_i^*(\theta)}{\partial^2 \theta}, \quad \widehat{h}_3 = \frac{\partial^3 h_i^*(\theta)}{\partial^3 \theta}, \quad \widehat{h}_4 = \frac{\partial^4 h_i^*(\theta)}{\partial^4 \theta},$$

with $\theta = \widehat{\theta}_i$ and $\boldsymbol{\gamma} = (\boldsymbol{a}^\top, \boldsymbol{b}^\top)^\top$. For each subject i, the gradient of the approximated marginal likelihood function with respect to $\boldsymbol{\gamma}$ can be represented as

$$\frac{\partial \log L_i^A(\boldsymbol{Y}_i|\theta; \boldsymbol{\Gamma}, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\gamma}} = -\frac{1}{2\widehat{h}_2} \cdot \frac{\partial \widehat{h}_2}{\partial \boldsymbol{\gamma}} - \frac{\partial \widehat{h}}{\partial \boldsymbol{\gamma}} + \frac{\partial \log(1+\tau)}{\partial \boldsymbol{\gamma}}, \tag{3.20}$$

with

$$\tau = -\frac{\widehat{h}_4}{8\widehat{h}_2^2} - \frac{5\widehat{h}_3^2}{24\widehat{h}_2^3}.$$

Each term in (3.20) can be achieved by plugging in the required derivative to it. Therefore, we present all derivatives in the following part. For $j = 1, \ldots, J$, we have

$$\frac{\partial \widehat{h}}{\partial a_j} = -Y_{ij}\mathrm{e}^{-a_j\widehat{\theta}_i-b_j}\cdot\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)\left(\frac{\widehat{\theta}_i\mathrm{e}^{a_j\widehat{\theta}_i+b_j}}{\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1}-\frac{\widehat{\theta}_i\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^2}\right) + \frac{\widehat{\theta}_i\cdot\left(1-Y_{ij}\right)\mathrm{e}^{a_j\widehat{\theta}_i+b_j}}{\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1}, \tag{3.21}$$

$$\frac{\partial \widehat{h}_2}{\partial a_j} = -\frac{2a_j\widehat{\theta}_i^2\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}}{\left(\mathrm{e_j}^{a_j\widehat{\theta}_i+b_j}+1\right)^3}+\frac{a_j\widehat{\theta}_i^2\mathrm{e}^{a_j\widehat{\theta}_i+b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^2}+\frac{2a_j\mathrm{e}^{a_j\widehat{\theta}_i+b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^2}, \tag{3.22}$$

$$\frac{\partial \widehat{h}_3}{\partial a_j} = -\frac{a_j\widehat{\theta}_i^3\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^3}+\frac{3a_j\widehat{\theta}_i^3\cdot\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}-1\right)\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^4} -\frac{a_j\widehat{\theta}_i^3\cdot\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}-1\right)\mathrm{e}^{a_j\widehat{\theta}_i+b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^3}-\frac{3a_j^2\cdot\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}-1\right)\mathrm{e}^{a_j\widehat{\theta}_i+b_j}}{\left(\mathrm{e}^{\widehat{\theta}_i a_j+b_j}+1\right)^3}, \tag{3.23}$$

and

$$\begin{aligned}\frac{\partial \widehat{h}_4}{\partial a_j} = &-\frac{4a_j\widehat{\theta}_i^4\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}\cdot\left(\mathrm{e_j}^{2a_j\widehat{\theta}_i+2b_j}-4\mathrm{e_j}^{a_j\widehat{\theta}_i+b_j}+1\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^5}\\ &+\frac{a_j^4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}\cdot\left(2\widehat{\theta}_i\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}-4\widehat{\theta}_i\mathrm{e}^{a_j\widehat{\theta}_i+b_j}\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^4}\\ &+\frac{a_j\widehat{\theta}_i^4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}\cdot\left(\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}-4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^4}\\ &+\frac{4a_j^3\mathrm{e}^{a_j\widehat{\theta}_i+b_j}\cdot\left(\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}-4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^4},\end{aligned} \tag{3.24}$$

By plugging (3.23) and (3.24), $\frac{\partial \log(1+\tau)}{\partial a_j}$ can be obtained as

$$\begin{aligned}&\frac{\partial \log(1+\tau)}{\partial a_j} \\ &= \left(1 - \frac{\widehat{h}_4}{8\widehat{h}_2^2} + \frac{5\widehat{h}_3^2}{\widehat{h}_2^3}\right)^{-1} \left[-\frac{8\frac{\partial \widehat{h}_4}{\partial a_j}\widehat{h}_2^2 - 16\frac{\partial \widehat{h}_2}{\partial a_j}\widehat{h}_4\widehat{h}_2}{(8\widehat{h}_2^2)^2} + \frac{240\frac{\partial \widehat{h}_3}{\partial a_j}\widehat{h}_2^3\widehat{h}_3 - 360\frac{\partial \widehat{h}_2}{\partial a_j}\widehat{h}_3^2\widehat{h}_2^2}{(24\widehat{h}_2^3)^2} \right].\end{aligned} \tag{3.25}$$

Similar derivations are also required for $\boldsymbol{b}$. For each subject i, the needed derivatives with respect to b_j are provided as below:

$$\frac{\partial \widehat{h}}{\partial b_j} = -\frac{\left(\mathrm{e}^{a_j\widehat{\theta}_i}Y_{ij} - \mathrm{e}^{a_j\widehat{\theta}_i}\right)\mathrm{e}^{b_j} + y}{\mathrm{e}^{a_j\widehat{\theta}_i + b_j +} + 1}, \tag{3.26}$$

$$\frac{\partial \widehat{h}_2}{\partial b_j} = \frac{a_j^2\mathrm{e}^{a_j\widehat{\theta}_i + b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i + b_j} + 1\right)^2} - \frac{2a_j^2\mathrm{e}^{2a_j\widehat{\theta}_i + 2b_j}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i + b_j} + 1\right)^3}, \tag{3.27}$$

and

$$\begin{aligned}&\frac{\partial \log(1+\tau)}{\partial b_j} \\ &= \left(1 - \frac{\widehat{h}_4}{8\widehat{h}_2^2} + \frac{5\widehat{h}_3^2}{\widehat{h}_2^3}\right)^{-1} \left[-\frac{8\frac{\partial \widehat{h}_4}{\partial b_j}\widehat{h}_2^2 - 16\frac{\partial \widehat{h}_2}{\partial b_j}\widehat{h}_4\widehat{h}_2}{(8\widehat{h}_2^2)^2} + \frac{240\frac{\partial \widehat{h}_3}{\partial b_j}\widehat{h}_2^3\widehat{h}_3 - 360\frac{\partial \widehat{h}_2}{\partial b_j}\widehat{h}_3^2\widehat{h}_2^2}{(24\widehat{h}_2^3)^2} \right],\end{aligned} \tag{3.28}$$

where

$$\frac{\partial \widehat{h}_3}{\partial b_j} = \frac{a_j^3\mathrm{e}^{a_j\widehat{\theta}_i + b_j} \cdot \left(\mathrm{e}^{2a_j\widehat{\theta}_i + 2b_j} - 4\mathrm{e}^{a_j\widehat{\theta}_i + b_j} + 1\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i + b_j} + 1\right)^4} \tag{3.29}$$

and

$$\begin{aligned}\frac{\partial \widehat{h}_4}{\partial b_j} &= \frac{a_j^4 \mathrm{e}^{a_j\widehat{\theta}_i+b_j}\cdot\left(2\mathrm{e}^{2a_j\widehat{\theta}_i}+2b_j-4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^4} \\ &\quad -\frac{4a_j^4\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}\cdot\left(\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}-4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^5} \\ &\quad +\frac{a_j^4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}\cdot\left(\mathrm{e}^{2a_j\widehat{\theta}_i+2b_j}-4\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^4}.\end{aligned} \tag{3.30}$$

For the gradient with respect to the variance σ^2 described in (3.1), we use the following explicit forms. To avoid negative values for σ^2 during optimization, we focus on updating the standard error σ instead. Then, we have

$$\frac{\partial \widehat{h}}{\partial \sigma} = \frac{1}{\sigma}-\frac{\left(\widehat{\theta}_i-\boldsymbol{x}_i\boldsymbol{\beta}\right)^2}{\sigma^3}, \tag{3.31}$$

$$\frac{\partial \widehat{h}_2}{\partial \sigma} = -\frac{2}{\sigma^3}, \tag{3.32}$$

$$\frac{\partial \widehat{h}_3}{\partial \sigma} = \frac{\partial \widehat{h}_4}{\partial \sigma}=0.$$

All terms in (3.20) have now been expressed in explicit forms with respect to $\boldsymbol{a}$ from (3.21) to (3.25), $\boldsymbol{b}$ from (3.26) to (3.30) and σ from (3.31) to (3.32). Additionally, we need the following derivatives to complete the gradient based on (3.33)

$$\frac{\partial^2 \widehat{h}_i}{\partial\theta\partial a_j}=-\frac{\left(\mathrm{e}^{2b_j}Y_{ij}-\mathrm{e}^{2b_j}\right)\mathrm{e}^{2a_j\widehat{\theta}_i}+\left(-\mathrm{e}^{b_j}a_j\widehat{\theta}_i+2\mathrm{e}^{b_j}Y_{ij}-\mathrm{e}^{b_j}\right)\mathrm{e}^{a_j\widehat{\theta}_i}+Y_{ij}}{\left(\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1\right)^2},$$

$$\frac{\partial^2 \widehat{h}_i}{\partial\theta\partial b_j}=-\frac{a_j\cdot\left[\left(\mathrm{e}^{b_j}Y_{ij}-\mathrm{e}^{b_j}\right)\mathrm{e}^{a_j\widehat{\theta}_i}+Y_{ij}\right]}{\mathrm{e}^{a_j\widehat{\theta}_i+b_j}+1},$$

and

$$\frac{\partial^2 \widehat{h}_i}{\partial\theta\partial\sigma} = -\frac{2\left(\widehat{\theta}_i - \boldsymbol{x}_i\boldsymbol{\beta}\right)}{\sigma^3}.$$

3.3 Computational Algorithm

In order to solve (3.13), the Newton-Raphson algorithm is not an appropriate choice because such an optimization algorithm requires the calculation of the inverse of the Hessian matrix. However, in our case, obtaining the exact form of the inverse Hessian matrix with respect to $\boldsymbol{\beta}$ is exceptionally challenging. As a result, we implement quasi-Newton methods to find the optimizer, as these methods have the capability to approximate the inverse Hessian matrix using the gradient during the computation process. This enables us to effectively operate the optimization process without using an exact inverse Hessian matrix.

Let $\boldsymbol{\Omega}$ denote the parameter vector, including item parameters and variance component. According to Andersson & Xin (2021), for each $\omega \in \boldsymbol{\Omega}$, the gradient with respect to parameter vector is written as

$$\boldsymbol{\nabla} = \sum_{i=1}^{N} \boldsymbol{\nabla}_i = \sum_{i=1}^{N} \frac{\partial \log L_i^A(\boldsymbol{\Omega}, \boldsymbol{\beta}|\boldsymbol{y}_i, \boldsymbol{x}_i)}{\partial\omega} + \frac{\partial\theta_i}{\partial\omega} \cdot \frac{\partial \log L_i^A(\boldsymbol{\Omega}, \boldsymbol{\beta}|\boldsymbol{y}_i, \boldsymbol{x}_i)}{\partial\theta}. \tag{3.33}$$

By employing a quasi-Newton method, we can represent the updating formula for the parameter vector at $r+1$th step as follows:

$$\boldsymbol{\Omega}^{(r+1)} = \boldsymbol{\Omega}^{(r)} - \lambda_s^{(r)} \boldsymbol{B}^{-1} \boldsymbol{\nabla}^{(r)}, \tag{3.34}$$

where $\boldsymbol{\Omega}^{(r)}$ represents the estimated parameter vector at rth step, $\lambda_s^{(r)}$ represents the step length used at rth step and $\boldsymbol{B}$ is the approximated Hessian matrix based on the gradient at rth step, $\boldsymbol{\nabla}^{(r)}$. Many literatures have studied the methods that

approximate the Hessian matrix based on the gradient vector obtained in (3.33). Berndt–Hall–Hall–Hausman (BHHH) (Berndt et al., 1974) algorithm is one of the widely used quasi-Newton methods. It substitutes the Hessian matrix with the outer product of the gradient. The BHHH algorithm relies on the information matrix equality, which is an approximation of the expected Hessian matrix. Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) is another frequently utilized quasi-Newton method in practice. The BFGS algorithm is an iterative procedure that aims to find the optimum value of an objective function. It gradually improves an approximation of the Hessian matrix by iteratively updating the gradient via a generalized secant method. According to the experience shared by Andersson & Xin (2021), it has been observed that the BFGS approximation achieves better convergence rates for small sample sizes. However, it takes a significantly higher number of iterations to converge in comparison to the BHHH approximation. Therefore, it is recommended to use BFGS if the sample size is small, whereas BHHH is more suitable for larger sample sizes.

The detailed optimization procedure of the proposed method can be summarized below:

- **Step 1**: Initialize item parameters $\boldsymbol{\Gamma}$, variance σ^2, regression coefficients $\boldsymbol{\beta}$ and latent trait θ_i for $i = 1, \ldots, N$ with some feasible values. And the initial values are denoted as $\boldsymbol{\Gamma}^{(0)}$, ${\sigma^2}^{(0)}$, $\boldsymbol{\beta}^{(0)}$ and ${\theta_i}^{(0)}$ respectively. ${\theta_i}^{(0)}$ is obtained by fitting an IRT model without considering any predictors in *ltm* R package. We choose constant vectors for $\boldsymbol{\Gamma}^{(0)}$ and $\boldsymbol{\beta}^{(0)}$, such vectors with all 1's. For ${\sigma^2}^{(0)}$, we use 1 as the initial value.

- **Step 2**: update the latent trait ${\theta_i}^{(r)}$ for $i = 1, \ldots, N$ by minimizing

$$h_i(\theta) = -\log\left[P(\boldsymbol{y_i}|\theta; \boldsymbol{\Gamma})\phi(\theta; \boldsymbol{x_i}^\top \boldsymbol{\beta}, \sigma^2)\right].$$

Specifically, we use BFGS algorithm for the optimization of the latent trait.

- **Step 3**: update $\mathbf{\Gamma}^{(r)}$ and $\sigma^{2(r)}$ using BHHH algorithm based on (3.34)
- **Step 4**: update $\boldsymbol{\beta}^{(r)}$ using group LASSO
- **Step 5**: When the largest difference between all parameters in step r and step $r-1$ is smaller than a threshold value, we stop the iterations and consider it as convergence. Otherwise, let $r = r+1$ and repeat Steps 2 to 4 until the convergence criterion mentioned above is met. In the simulation, we use 0.00001 as our threshold value.

In Step 4 above, we employ a fast unified algorithm, proposed by Yang & Zou (2015), to solve group LASSO problems.

3.4 Selection of Tuning Parameter λ

The tuning parameter λ in (3.13) plays an essential role in selecting important groups of covariates because it determines the strength of the penalty applied to the variables. When λ grows bigger, it shrinks more coefficients towards zero. Conversely, as λ decreases, fewer coefficients are shrunk, and more variables are included in the model. Therefore, an appropriate value of λ is crucial for our proposed variable selection method.

As mentioned in Section 1.2.1.9, GCV is a shortcut of LOOCV, and it is more computationally efficient than the traditional cross-validation approach. We propose to use the GCV style statistic in our method, which is defined by

$$GCV(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\widehat{\theta}_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}}{1 - p(\lambda)/N}\right)^2, \tag{3.35}$$

where $\widehat{\theta}_i$ and $\widehat{\boldsymbol{\beta}}$ are estimated from our method, and $p(\lambda)$ represents the effective degrees of freedom. Let $\boldsymbol{W}_1 = \text{diag}(|\widehat{\beta}_j|)$, $\boldsymbol{W}_2 = \text{diag}(2|\widehat{\beta}_j|)$, and $\boldsymbol{W}_1^-$ and $\boldsymbol{W}_2^-$ denote a generalized inverse matrix of $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ respectively. We consider four versions of $p(\lambda)$ in (3.35) :

- GCV1: $p(\lambda) = \text{trace}\{\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{W}_1^-)^{-1}\boldsymbol{X}^\top\}$
- GCV2: $p(\lambda) =$ the number of non-zero coefficients in $\widehat{\boldsymbol{\beta}}$.
- GCV3: $p(\lambda) = \text{trace}\{\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{W}_1^-)^{-1}\boldsymbol{X}^\top\}-$ number of zero coefficients in $\widehat{\boldsymbol{\beta}}$.
- GCV4: $p(\lambda) = \text{trace}\{\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda\boldsymbol{W}_2^-)^{-1}\boldsymbol{X}^\top\}-$ number of zero coefficients in $\widehat{\boldsymbol{\beta}}$.

We select the best λ that results in the smallest value of GCV defined in (3.35). In Section 3.5.2, we will utilize those four different versions of GCV and compare the selection performance based on them.

3.5 Simulation Studies

3.5.1 Setting

To examine the performance of the proposed method, we perform simulation studies with several various settings. We also compare results with the two-step approach mentioned in Section 1.5.2. Specifically, in the first step, we fit a unidimensional IRT model to estimate item parameters and the latent trait without considering a set of covariates, and in the second step, we use either stepwise selection or LASSO for variable selection. To estimate parameters in the first step, we implement *ltm* package in R (Rizopoulos, 2006). This package uses the Gauss-Hermite quadrature rule to estimate the integral in the marginal likelihood function. In the simulation studies, we mainly focus on a 2PL model that has been mentioned in Section 1.5.1.2.

The 2PL model contains two item parameters: discrimination ($\boldsymbol{a}$) and easiness ($\boldsymbol{b}$). The responses to test items are binary, taking values of either 1 (indicating a correct response) or 0 (indicating a wrong response). Moreover, We use the following several measurements to measure the performance of each approach: the number of zero coefficients that are correctly estimated as zero (denoted as 'CZ'), the number of non-zero coefficients that are incorrectly set to zero (denoted as 'IZ'), the number of unimportant groups which are correctly estimated as unimportant group (denoted as 'CZG'), the number of important groups which are incorrectly set to zero (denoted as 'IZG'), the frequency of selecting the correct model (denoted as 'C'), the frequency of over-selecting models (denoted as 'O') and the frequency of under-selecting models (denoted as 'U'). To measure the performance of estimation on latent regression coefficients, we apply the root mean squared error (RMSE) defined as

$$\text{RMSE} = \frac{\sum_{i=1}^{I} \sqrt{||\boldsymbol{\beta}^{(i)} - \boldsymbol{\beta}||^2/p}}{I},$$

where $\boldsymbol{\beta}^{(i)}$ represents the estimated coefficients vector at ith replication, $\boldsymbol{\beta}$ represents the true values of coefficients, p represents the number of covariates and I is the number of total replications. We collect $I = 100$ observations for each simulation setting, and the medians are recorded.

- **Example 1:** In this example, we mimic the setting with the NAEP sample data that will be used in Section 3.6. There are $N = 2000$ subjects and $J = 15$ question items. The set of covariates follows a similar distribution to the NAEP sample data. For the latent regression coefficients, we set

$$\begin{aligned}\boldsymbol{\beta} =&(0, -0.8, -0.3, 0.2, -0.5, -0.2, 0.6, -0.6, -0.1, 0.2, 0.3, -0.1, -0.2,\\ &0.1, -0.3, -0.2, -0.7, 0, 0, 0, 0.1, 0.2, 0.2, 0.3, 0, 0, 0, 0)^\top\end{aligned}$$

There are nine categorical variables in total, and the group size for each variable, that is, the number of dummy variables converted from each variable, is $1, 5, 2, 4, 1, 4, 3, 4$, and 4. Specifically, the 1st, 7th, and the last group are unimportant factors. Additionally, we generate item parameters $\boldsymbol{a}$ from a uniform distribution between 0.3 and 2 and $\boldsymbol{b}$ from a uniform distribution between -2 and 2.

- **Example 2:** 6 variables $Z_1, \ldots, Z_6$ are generated by a multivariate normal distribution with mean zeros and covariance between Z_i and Z_j being $0.5^{|i-j|}$. Then, we created 6 groups of categorical variables corresponding to $Z_1, \ldots, Z_6$ respectively. Then, Z_1 is categorized as $0, 1, 2, 3, 4$ if it is less than $\Phi^{-1}(\frac{1}{5})$, between $\Phi^{-1}(\frac{1}{5})$ and $\Phi^{-1}(\frac{2}{5})$, between $\Phi^{-1}(\frac{2}{5})$ and $\Phi^{-1}(\frac{3}{5})$, between $\Phi^{-1}(\frac{3}{5})$ and $\Phi^{-1}(\frac{4}{5})$ or greater than $\Phi^{-1}(\frac{4}{5})$. Z_2 and Z_3 are categorized as $0, 1, 2, 3$ if it is less than $\Phi^{-1}(\frac{1}{4})$, between $\Phi^{-1}(\frac{1}{4})$ and $\Phi^{-1}(\frac{1}{2})$, between $\Phi^{-1}(\frac{1}{2})$ and $\Phi^{-1}(\frac{3}{4})$, or greater than $\Phi^{-1}(\frac{3}{4})$. Z_4, Z_5 and Z_6 are categorized as 0,1,2 if it is less than $\Phi^{-1}(\frac{1}{3})$, greater than $\Phi^{-1}(\frac{2}{3})$ or in between. The latent trait for each subject i was generated from

$$\begin{aligned}\theta_i =& I(Z_{1i} = 0) - I(Z_{1i} = 1) - I(Z_{1i} = 3) + I(Z_{1i} = 4) \\ &+ I(Z_{2i} = 0) - 0.5I(Z_{2i} = 1) - 0.5I(Z_{2i} = 3) \\ &+ 0.5I(Z_{3i} = 0) + 0.5I(Z_{3i} = 1) - (Z_{3i} = 3) \\ &+ I(Z_{4i} = 0) - I(Z_{4i} = 2) + \epsilon_i,\end{aligned}$$

where $I(\cdot)$ is the indicator function. The error term ϵ_i is assumed to follow a normal distribution with mean 0 and variance 1. Then, we simulated a test response data with

  - *Case 1*: $N = 500$ test takers and $J = 20$ test items

- *Case 2*: $N = 1000$ test takers and $J = 40$ test items

with item parameters

$$\boldsymbol{b} \sim Uniform(-2, 2),$$

$$\boldsymbol{a} = (1, 1, 1, 1, 1, 0.8, 0.8, 0.8, 0.8, 0.8, 0.4, 0.4, 0.4, 0.4, 0.4, 0.2, 0.2, 0.2, 0.2, 0.2),$$

and a vector of latent trait $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top$ for $m = 1, \ldots, 500$.

- **Example 3:** Same setup as example 2 except the latent trait for the subject i is now given by:

$$\begin{aligned}\theta_i =& 0.7I(Z_{1i} = 0) - 0.7I(Z_{1i} = 1) - 0.5I(Z_{1i} = 3) + 0.5I(Z_{1i} = 4) \\ &+ 0.4I(Z_{2i} = 0) - 0.1I(Z_{2i} = 1) - 0.1I(Z_{2i} = 3) \\ &+ 0.2I(Z_{3i} = 0) + 0.1I(Z_{3i} = 1) - 0.6I(Z_{3i} = 3) \\ &+ 0.3I(Z_{4i} = 0) - 0.3I(Z_{4i} = 2) + \epsilon_i\end{aligned}$$

We also consider two cases:

- *Case 1*: $N = 500$ test takers and $J = 20$ test items
- *Case 2*: $N = 1000$ test takers and $J = 40$ test items

- **Example 4:** 8 variables $Z_1, \ldots, Z_8$ are generated by a multivariate normal distribution with mean zeros and covariance between Z_i and Z_j being $0.5^{|i-j|}$. Then we created 8 groups of categorical variables corresponding to $Z_1, \ldots, Z_8$ respectively. Then, Z_1 is categorized as $0, 1, 2, 3, 4, 5$ if it is less than $\Phi^{-1}(\frac{1}{6})$, between $\Phi^{-1}(\frac{1}{6})$ and $\Phi^{-1}(\frac{2}{6})$, between $\Phi^{-1}(\frac{2}{6})$ and $\Phi^{-1}(\frac{3}{6})$, between $\Phi^{-1}(\frac{3}{6})$ and $\Phi^{-1}(\frac{4}{6})$,between $\Phi^{-1}(\frac{4}{6})$ and $\Phi^{-1}(\frac{5}{6})$ or greater than $\Phi^{-1}(\frac{5}{6})$. Z_2 is categorized as $0, 1, 2, 3, 4$ if it is less than $\Phi^{-1}(\frac{1}{5})$, between $\Phi^{-1}(\frac{1}{5})$ and $\Phi^{-1}(\frac{2}{5})$, between $\Phi^{-1}(\frac{2}{5})$ and $\Phi^{-1}(\frac{3}{5})$, between $\Phi^{-1}(\frac{3}{5})$ and $\Phi^{-1}(\frac{4}{5})$or greater than $\Phi^{-1}(\frac{4}{5})$. Z_3,

Z_4 and Z_5 are categorized as $0, 1, 2, 3, 4$ if they are less than $\Phi^{-1}(\frac{1}{4})$, between $\Phi^{-1}(\frac{1}{4})$ and $\Phi^{-1}(\frac{1}{2})$, between $\Phi^{-1}(\frac{1}{2})$ and $\Phi^{-1}(\frac{3}{4})$, or greater than $\Phi^{-1}(\frac{3}{4})$. Z_6 and Z_7 are categorized as 0,1,2 if it is less than $\Phi^{-1}(\frac{1}{3})$, greater than $\Phi^{-1}(\frac{2}{3})$ or in between. Z_8 is categorized as either 0 or 1 if it is either less than or greater than $\Phi^{-1}(\frac{1}{2})$.

Latent trait for each subject i was generated from:

$$\begin{aligned}\theta_i = {} & 0.5I(Z_{1i} = 0) - 0.5I(Z_{1i} = 1) - 0.5I(Z_{1i} = 3) + 0.5I(Z_{1i} = 4) \\ & - 0.5I(Z_{1i} = 5) + 0.5I(Z_{1i} = 6) \\ & + 0.5I(Z_{2i} = 0) + 0.1I(Z_{2i} = 1) - 0.1I(Z_{2i} = 3) + 0.1(Z_{2i} = 4) \\ & - 0.2(Z_{2i} = 5) \\ & + 0.2I(Z_{3i} = 0) - 0.4I(Z_{3i} = 1) + 0.1(Z_{3i} = 3) - 0.4I(Z_{3i} = 4) \\ & + 0.1I(Z_{4i} = 0) - 0.5I(Z_{4i} = 2) - 0.3I(Z_{4i} = 3) + 0.1I(Z_{4i} = 4) + \epsilon_i,\end{aligned}$$

where $I(\cdot)$ is the indicator function. The error term ϵ_i is assumed to follow a normal distribution with mean 0 and variance 1. Then, we simulated a test response data with

- *Case 1*: $N = 500$ test takers and $J = 20$ test items
- *Case 2*: $N = 1000$ test takers and $J = 40$ test items

with item parameters

$$\boldsymbol{b} \sim Uniform(-2, 2),$$

$$\boldsymbol{a} = (1, 1, 1, 1, 1, 0.8, 0.8, 0.8, 0.8, 0.8, 0.4, 0.4, 0.4, 0.4, 0.4, 0.2, 0.2, 0.2, 0.2, 0.2),$$

and a vector of latent trait $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^\top, m = 1, \ldots, 500$.

Table 3.1: Grouped variable selection results of Example 1

N	J	Method	CZ	IZ	CZG	IZG	C	O	U	RMSE
500	15	Proposed (GCV1)	8	0	3	0	26	40	34	0.2193
		Proposed (GCV2)	8	0	3	0	29	35	36	0.2185
		Proposed (GCV3)	8	0	3	0	29	35	36	0.2185
		Proposed (GCV4)	8	0	3	0	29	35	36	0.2185
		2-step (stepwise)	8	3	5	2	0	0	100	0.4503
		2-step (LASSO)	3	0	0	0	0	10	86	0.3558
2000	15	Proposed (GCV1)	8	0	3	0	33	34	33	0.1722
		Proposed (GCV2)	8	0	3	0	34	34	32	0.1724
		Proposed (GCV3)	8	0	3	0	34	34	32	0.1724
		Proposed (GCV4)	8	0	3	0	34	34	32	0.1724
		2-step (stepwise)	8	3	5	2	3	0	97	0.4398
		2-step (LASSO)	3	0	0	0	1	5	94	0.3672
		Oracle	8	0	6	0	100	0	0	

3.5.2 Results

We present the simulation results using the measurements as mentioned earlier in Tables 3.1 to 3.4. Additionally, we include the performance of the two-step approaches, which utilize stepwise selection based on BIC and LASSO, separately. We assess the variable selection performance at both the group and individual levels.

Table 3.1 reports the performance of the proposed method and the other 2-step approaches. The case $N = 2,000$ and $J = 15$ mimics the applied real data, and the proposed method outperforms the other two 2-step approaches. In particular, the proposed method has better chances of selecting the true models and consistently smaller RMSEs compared to the other approaches. The case $N = 500$ and $J = 20$ is a more extreme case in the real world when there is a smaller number of examinees in

Table 3.2: Grouped variable selection results for case 1 and case 2 of Example 2

N	J	Method	CZ	IZ	CZG	IZG	C	O	U	RMSE
500	20	Proposed (GCV1)	4	0	2	0	97	3	0	0.4138
		Proposed (GCV2)	4	0	2	0	98	2	0	0.4140
		Proposed (GCV3)	4	0	2	0	98	2	0	0.4140
		Proposed (GCV4)	4	0	2	0	98	2	0	0.4140
		2-step (stepwise)	4	0	2	0	100	0	0	0.4353
		2-step (LASSO)	1	0	0	0	1	99	0	0.9076
1000	40	Proposed (GCV1)	4	0	2	0	81	19	0	0.2301
		Proposed (GCV2)	4	0	2	0	85	15	0	0.2201
		Proposed (GCV3)	4	0	2	0	85	15	0	0.2201
		Proposed (GCV4)	4	0	2	0	85	15	0	0.2201
		2-step (stepwise)	4	0	2	0	100	0	0	0.3924
		2-step (LASSO)	1	0	0	0	8	92	0	0.9328
		Oracle	4	0	2	0	100	0	0	

the assessment. The proposed method has outperformed the other two approaches on both variable selection and coefficient estimation.

Table 3.2 considers an ideal case where all coefficients are relatively large, and all methods perform well on variable selection. The two-step stepwise approach ideally selects all correct variables; however, as introduced in the previous section, two-step approaches are biased; hence they have larger RMSEs in coefficient estimation than the proposed method.

Table 3.3 considers the case where the number of groups and group size stays the same with Table 3.2, but coefficients vary. The proposed method outperforms both two-step approaches in a small sample size case of $N = 500$, and the performance of the proposed method and two-step stepwise selection are comparable in a large sample size case of $N = 1000$.

Table 3.3: Grouped variable selection results for case 1 and case 2 of Example 3

N	J	Method	CZ	IZ	CZG	IZG	C	O	U	RMSE
500	20	Proposed (GCV1)	4	0	2	0	54	26	20	0.1521
		Proposed (GCV2)	4	0	2	0	58	20	22	0.1503
		Proposed (GCV3)	4	0	2	0	58	20	22	0.1503
		Proposed (GCV4)	4	0	2	0	58	20	22	0.1503
		2-step (stepwise)	4	2	2	1	20	1	79	0.1778
		2-step (LASSO)	2	1	0	0	0	26	74	0.5012
1000	40	Proposed (GCV1)	4	0	2	0	80	7	13	0.1403
		Proposed (GCV2)	4	0	2	0	81	10	9	0.1385
		Proposed (GCV3)	4	0	2	0	81	10	9	0.1385
		Proposed (GCV4)	4	0	2	0	81	10	9	0.1385
		2-step (stepwise)	4	0	2	0	82	0	18	0.1325
		2-step (LASSO)	1	0	0	0	1	51	48	0.5186
		Oracle	4	0	2	0	100	0	0	

Table 3.4 considers the case where both the number of groups and group size increase and coefficients vary. The proposed method outperforms both two-step approaches on variable selection and coefficient estimation for both small and large sample sizes.

In conclusion, the two-step stepwise approach performs well under ideal cases where all coefficients are large enough or when the sample size is large. The proposed method outperforms two-step approaches under more sophisticated cases, which are commonly seen in practice.

Table 3.4: Grouped variable selection results for case 1 and case 2 of Example 4

N	J	Method	CZ	IZ	CZG	IZG	C	O	U	RMSE
500	20	Proposed (GCV1)	10	0	3	0	42	16	42	0.1519
		Proposed (GCV2)	10	0	3	0	44	12	44	0.1501
		Proposed (GCV3)	10	0	3	0	44	12	44	0.1501
		Proposed (GCV4)	10	0	3	0	44	12	44	0.1501
		2-step (stepwise)	12	9	4	2	2	0	98	0.1878
		2-step (LASSO)	5	0	2	0	0	11	89	0.3810
1000	40	Proposed (GCV1)	12	0	4	0	65	9	22	0.1201
		Proposed (GCV2)	12	0	4	0	70	11	18	0.1183
		Proposed (GCV3)	12	0	4	0	70	11	18	0.1183
		Proposed (GCV4)	12	0	4	0	70	11	18	0.1183
		2-step (stepwise)	12	0	4	0	42	0	58	0.1298
		2-step (LASSO)	5	0	0	0	2	16	72	0.3949
		Oracle	12	0	4	0	100	0	0	

3.6 Real Data Example

3.6.1 Description

We have implemented the proposed method for selecting grouped variables in the datasets of the National Assessment of Educational Progress (NAEP) from 2005. NAEP is an ongoing assessment program across the United States that evaluates students' academic performance in diverse subjects such as reading, mathematics, science, and more. In this real data example, we focus on selecting several person predictors that significantly affect the mathematics ability of a student.

We use a sample data set downloaded from R package 'EdSurvey' (Bailey et al., 2023). The data set contains assessment responses on 17 items from 2348 8th-grade students. We use those responses as our $\boldsymbol{Y}$. Additionally, we are able to extract

several person predictors for each student to construct a design matrix $\boldsymbol{X}$. In particular, x_1 is gender (0= white, 1= black, 2= Hispanic, 3= Asian/pacific islander), 4= American Indian/ Alaska Native, 5= other), x_2 is gender (0=male, 1=female), x_3 is if the student classified as an English language learner (ELL) (0=yes, 1=no, 2= formerly ELL), x_4 is parental education level (0=did not finish high school, 1= graduated from high school, 2=some education after high school, 3=graduated college, 4=I don't know), x_5 is if students have access to a computer at home (0=yes, 1=no), x_6 is the number of days absent from school last month (0=none, 1= 1 to 2 days, 2= 3 to 4 days, 3= 5 to 10 days, 4= more than 10 days), x_7 is if the student speaks a language other than English at home (0=never, 1=once in a while,2= half the time, 3=all or most of the time), x_8 is how often the student talks about studies at home (0=never or hardly ever, 2=once every few weeks, 3= about once a week, 4= 2 or 3 times a week, 5=every day), x_9 is how often the student uses computers to make up tests for individual students (0=never use computers, 1=sometimes computers, 2=always use computers), x_{10} is how often the student uses computers for individual tests (0=never use computers, 1=sometimes use computers, 2=always use computers), x_{11} is how often the student uses math tool for math concepts as a computer activity (0=never or hardly ever, 1=once or twice per month, 2=once or twice a week, 3=almost every day), x_{12} is how often the student develops math curricula and assignments as a computer activity (0=never or hardly ever, 1=once or twice/month, 2=once or twice a week, 3=almost every day), x_{13} is how often the student uses a gradebook program as a computer activity (0=never or hardly ever, 1=once or twice/month, 2=once or twice a week, 3=almost every day), x_{14} is how often the student posts homework, schedule information as a computer activity (0=never or hardly ever, 1=once or twice/month, 2=once or twice a week, 3=almost every day). In total, there are 14 categorical predictors. Since x_{11}, x_{12},

x_{13}, and x_{14} all measure the frequency of performing computer activities, we treat those as a group. Similarly, for x_9 and x_{10}, they both represent computer access for students' tests, so we also treat them as a group of predictors. Therefore, 10 groups of predictors are entered into our method for variable selection.

3.6.2 Results

After applying our method to the NAEP sample data, we have successfully identified several significant variables that related to the mathematics skill of those eighth grade students. The selection results are listed in Table 3.5. We also have observed some variables' coefficients are shrunk to zero. In this case, we consider those variables as unimportant, and they are listed in Table 3.6. Besides the proposed method, we employ a two-step approach by using stepwise selection based on BIC. The selection results are demonstrated in Table 3.9 and 3.10. Furthermore, we implement stepwise selection by using plausible values provided in NAEP data as the response variable. The selection results are presented in Table 3.7 and 3.8.

Based on the selection results of person predictors, we can see that our method successfully selects six important predictors related to students' math skills: students' race, their parent's education level, English Language Learner (ELL) status, if they have computer access at home, how frequently they were absent last month and how often students talk about study at home. The results of the estimated coefficients of these predictors provide additional insights. Among different racial groups, white students exhibit better mathematical abilities compared to other races, except for Asian or Pacific Islander students, who have slightly higher math skills, although the difference is relatively small. The math ability of students is also significantly influenced by their ELL status. Non-ELL students and those who were formerly ELL demonstrate considerably better skills.

Table 3.5: Important predictors to mathematics ability of 8th-grade students and their estimated coefficients from the proposed approach

Variable	Coefficient
Race (White) - ref.	-
Race (Black)	-0.8796
Race (Hispanic)	-0.5435
Race (Asian/Pacific Islander)	0.0686
Race (Amer Ind/ Alaska Native)	-0.1404
Race (Other)	-0.0402
English language learner (Yes) -ref.	-
English language learner (No)	0.7232
English language learner (Formerly)	0.7004
Parental education level (Did not finish H.S) - ref.	-
Parental education level (Graduated H.S)	-0.0771
Parental education level (Some ed after H.S)	0.1356
Parental education level (Graduated college)	0.3088
Parental education level (I don't know)	-0.2298
Computer at home (Yes) - ref.	-
Computer at home (No)	-0.3180
Days absent from school last month (None) - ref.	-
Days absent from school last month (1-2 days)	-0.0935
Days absent from school last month (3-4 days)	-0.3094
Days absent from school last month (5-10 days)	-0.2372
Days absent from school last month (More than 10 days)	-0.7268
Talk about study at home (Never) - ref.	-
Talk about study at home (Once every few weeks)	-0.0472
Talk about study at home (Once a week)	0.0484
Talk about study at home (2-3 times a week)	0.0823
Talk about study at home (Every day)	-0.0292

Table 3.6: Unimportant predictors to mathematics ability of 8th-grade students from the proposed approach

Variable
Gender
Language other than English spoken at home
Use computers to make up tests for individual student
Use computers for individual tests for all students
Computer activities: Math tool for math concepts
Computer activities: Develop math curricula,assignments
Computer activities: Use a gradebook program
Computer activities: Post homework, schedule info

Furthermore, students whose parents have graduated from college demonstrate the highest level of math ability. The presence of computer access at home is also positively associated with improved math skills. This finding highlights the significance of accessing computers for student learning outcomes. Moreover, the frequency of absence from last month negatively related to math skills. In other words, the more absences a student has, the lower their math ability tends to be. Finally, the last selected factor is how often the student talks about study at home. The coefficient results show that students who engage in frequent discussions about their studies at home tend to display better math skills.

There are several predictors have also been excluded from the final model. The selection results show that students' math skills are not significantly affected by the following predictors: students' gender, how often they speak a language other than English spoken at home, computer access for student tests, and engagement in multiple computer activities. For instance, irrespective of whether students speak a language other than English at home, it does not appear to influence their learning ability in mathematics.

In addition to using the proposed method, we also fit a linear model with plausible

values as response and the same factors as covariates, followed by stepwise selection. Plausible values are computed based on the posterior distribution of the latent trait and can be regarded as estimations of the distribution of the latent trait. In the sample NAEP data, six plausible values are available. Consequently, we compute the average of these plausible values as our response variable.

Table 3.7 presents the important factors selected using this method, and Table 3.8 lists the insignificant factors. It is shown that the chosen variables are consistent with those selected by the proposed method. In practice, the computation of plausible values might be complicated; thus, plausible values may not be easily accessible in real-world data.

Table 3.9 presents the selected factors using a two-step approach. In the first step, we obtain the latent trait estimates from the *ltm* package and then perform stepwise selection based on BIC. Compared to the proposed method and regression on plausible values, the naive approach ignores two factors: whether the students have access to a computer at home and how frequently they talk about study at home. Including these two variables makes more sense since they are closely related to students' learning attitudes and resources. In today's digital age, many learning materials and resources are accessible through computers, and having more access to computers could provide students with increased opportunities for studying.

In general, the two-step approach continues to exclude more variables in the final model, as we have seen in simulation studies. Additionally, the selection results of the proposed method align with the results obtained from modeling with plausible values. Therefore, our method can be highly beneficial when plausible values are unavailable in the assessment data.

Table 3.7: Important predictors to mathematics ability of 8th-grade students and their estimated coefficients from two-step approach based on plausible values

Variable	Coefficient
Race (White) - ref.	-
Race (Black)	-0.8113
Race (Hispanic)	-0.3591
Race (Asian/Pacific Islander)	0.1435
Race (Amer Ind/ Alaska Native)	-0.5282
Race (Other)	-0.3570
English language learner (Yes) -ref.	-
English language learner (No)	0.6932
English language learner (Formerly)	0.6121
Parental education level (Did not finish H.S) - ref.	-
Parental education level (Graduated H.S)	-0.0951
Parental education level (Some ed after H.S)	0.2894
Parental education level (Graduated college)	0.4132
Parental education level (I don't know)	-0.1364
Computer at home (Yes) - ref.	-
Computer at home (No)	-0.2044
Days absent from school last month (None) - ref.	-
Days absent from school last month (1-2 days)	-0.0935
Days absent from school last month (3-4 days)	-0.3094
Days absent from school last month (5-10 days)	-0.2372
Days absent from school last month (More than 10 days)	-0.7268
Talk about study at home (Never) - ref.	-
Talk about study at home (Once every few weeks)	-0.0054
Talk about study at home (Once a week)	0.2126
Talk about study at home (2-3 times a week)	0.2749
Talk about study at home (Every day)	0.0703

Table 3.8: Unimportant predictors to mathematics ability of 8th-grade students from two-step approach based on plausible values

Variable
Gender
Language other than English spoken at home
Use computers to make up tests for individual students
Use computers for individual tests for all students
Computer activities: Math tool for math concepts
Computer activities: Develop math curricula, assignments
Computer activities: Use a gradebook program
Computer activities: Post homework, schedule info

Table 3.9: Important predictors to mathematics ability of 8th-grade students and their estimated coefficients from a two-step approach based on estimation from *ltm* R package

Variable	Coefficient
Race (White) - ref.	-
Race (Black)	-0.5981
Race (Hispanic)	-0.2819
Race (Asian/Pacific Islander)	0.0982
Race (Amer Ind/ Alaska Native)	-0.3220
Race (Other)	-0.4034
English language learner (Yes) -ref.	-
English language learner (No)	0.5020
English language learner (Formerly)	0.4884
Parental education level (Did not finish H.S) - ref.	-
Parental education level (Graduated H.S)	-0.0807
Parental education level (Some ed after H.S)	0.2387
Parental education level (Graduated college)	0.3186
Parental education level (I don't know)	-0.1536
Days absent from school last month (None) - ref.	-
Days absent from school last month (1-2 days)	-0.0682
Days absent from school last month (3-4 days)	-0.2312
Days absent from school last month (5-10 days)	-0.1603
Days absent from school last month (More than 10 days)	-0.4936

Table 3.10: Unimportant predictors to mathematics ability of 8th-grade students from two-step approach based on estimation from *ltm* R package

Variable
Gender
Computer access at home
Talk about study at home
Language other than English spoken at home
Use computers to make up tests for individual student
Use computers for individual tests for all students
Computer activities: Math tool for math concepts
Computer activities: Develop math curricula,assigments
Computer activities: Use a gradebook program
Computer activities: Post homework, schedule information

Chapter 4

Discussion & Future work

4.1 Discussion

In this dissertation, we developed new variable selection methods for identifying grouped variables within linear mixed effects and latent variable models. Specifically, we construct frameworks for a penalized likelihood selection approach by integrating a group LASSO penalty into appropriate likelihood functions. In these two statistical modelings, grouped variables are frequently seen in practice. However, selection methods aimed at group-level selection remain largely unexplored. Therefore, our approaches address such gaps and make variable selection more efficient and accurate.

In simulation studies on the linear mixed models, the proposed method demonstrates superior performance in terms of selection accuracy compared to other methods. It is particularly effective when the sample size is small. It is known that longitudinal studies usually involve small sample sizes due to the cost of collecting observations. In this scenario, our method could be a great choice. In the simulation studies related to the IRT models, our proposed method presents a superior capability in detecting variables that have small coefficients but are significantly associated with the latent variable, while the two-step approach tends to ignore them. In addi-

tion, the proposed method results in a better estimation of model parameters than the two-step approach in both small and large samples. In practical applications, this method could effectively identify critical factors that impact the latent variables measured from test items.

We use two proposed methods on real education data. In the first topic, we use longitudinal data from ECLS-K that follows a group of students from Kindergarten through the 8th grade. We consider both continuous and categorical variables to enter the model and perform random effects selection. The first proposed method is able to select important random effects at the group level for grouped variables. In the second topic, we use a sample NAEP data that contains item responses from 8th-grade students in the U.S. We consider a set of candidate factors that might affect students' math skills and perform variable selection using the proposed method and the traditional two-step approach. The second proposed method outperforms the traditional two-step approach in estimating parameters and selection accuracy.

Nevertheless, there remain areas for potential improvement in these two projects. For the first project, we can keep working on the computation part to make computation more efficient. In the second topic, we can try more methods on tuning parameters, such as cross-validation, to make selection and parameter estimation even better. A suitable tuning parameter can closely affect the performance of parameter estimations and the selection of coefficients.

4.2 Future Work

There are potential extensions to my current research topics. In the linear mixed model, it is common to consider selecting both grouped random effects and fixed effects simultaneously. This is an important aspect to study in the future, as it will help to determine the most informative variables. Moreover, I focused on variable

selection in unidimensional item response theory models in this dissertation, which considers one latent trait in the model. Multidimensional item response theory models extend the unidimensional IRT models by accounting for the possibility that multiple latent traits or abilities may underlie the responses to test items. We can extend our selection idea to multidimensional IRT models. Additionally, it would be possible to explore latent trait selection for multidimensional IRT models in static and dynamic setups.

References

Ahn, M., Zhang, H. H., & Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, *22*(4), 1539-1562. doi: 10.5705/ss.2011.054

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, *60*, 255–265.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, *16*, 125-127. doi: 10.1080/00401706.1974.10489157

Andersson, B., & Xin, T. (2021). Estimation of latent regression item response theory models using a second-order Laplace approximation. *Journal of Educational and Behavioral Statistics*, *46*(2), 244–265. doi: 10.3102/1076998620945199

Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., ... Christensen, A. A. (2023). Edsurvey: Analysis of nces education survey and assessment data [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=EdSurvey` (R package version 3.1.0)

Berndt, E. K., Hall, B. H., Hall, R. E., & Hausman, J. A. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, *3*, 653–665.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In *Statistical Theories of Mental Test Scores* (chap. 17–20).

Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, *66*(4), 1069-1077. doi: 10.1111/j.1541-0420.2010.01391.x

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, *37*(4), 373-384. doi: 10.2307/1269730

Breiman, L., & Spector, P. (1992). Submodel selection and evaluation in regression. The X-random case. *International Statistical Review/Revue Internationale de Statistique*, *60*(3), 291-319. doi: 10.2307/1403680

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA Journal of Applied Mathematics*, *6*, 76-90. doi: 10.1093/imamat/6.1.76

Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement*, *52*, 200-222. doi: 10.1111/jedm.12072

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of The Royal Statistical Society Series A: Statistics in Society*, *158*, 419-444. doi: 10.2307/2983440

Chen, Y., Li, X., & Zhang, S. (2018). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, *84*, 124-146. doi: 10.1007/s11336-018-9646-5

Chen, Z., & Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, *59*, 762–769. doi: 10.1111/j.0006-341x.2003.00089.x

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, *31*, 377–403.

Derksen, S. A., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*, 265-282.

Early childhood longitudinal studies program - kindergarten class of 1998-99. (1998-2007). Retrieved from https://nces.ed.gov/ecls/kindergarten.asp

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348 - 1360. doi: 10.1198/016214501753382273

Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, *13*(3), 317-322. doi: 10.1093/comjnl/13.3.317

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of Computation*, *24*, 23-26. doi: 10.1090/S0025-5718-1970-0258249-6

Greven, S., & Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Communications in Statistics - Theory and Methods*, *97*(4), 773-789. doi: 10.1093/biomet/asq042

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Sage.

Harrell, L. (2015). *Analysis strategies for planned missing data in health sciences and education research* (Unpublished doctoral dissertation). University of California, Los Angeles. (Available at https://escholarship.org/uc/item/7cz9q2vm)

Hastie, T., Tibshirani, R., & Jerome, F. (2009). *The elements of statistical learning.* Springer Series in Statistics.

Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, *12*(1), 69–82. doi: 10.1080/00401706.1970.10488635

Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634

Holiday, D. B., Ballard, J. E., & Mckeown, B. C. (1995). PRESS-related statistics: Regression tools for cross-validation and case diagnostics. *Medicine and science in sports and exercise*, *27*(4), 612-620.

Hosmer Jr, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: John Wiley.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. doi: 10.1093/biomet/76.2.297

Hutchinson, T. P. (1991). *Ability, partial information, guessing: statistical modelling applied to multiple, choice tests.* Adelaide: Rumsby Scientific Publishing.

Knott, M., & Bartholomew, D. J. (1999). *Latent variable models and factor analysis* (2nd ed., Vol. 7). London, UK: Kendall's library of statistics.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963-974. doi: 10.2307/2529876

Leng, C., Lin, Y., , & Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, *16*(4), 1273-1284.

Li, Y., Wang, S., Song, P. X.-K., Wang, N., Zhou, L., & Zhu, J. (2018). Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Statistics and its Interface*, *11*(4), 721–737. doi: 10.4310/SII.2018.v11.n4.a15

Lin, Y., & Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, *34*, 2272-2297. doi: 10.1214/009053606000000722

Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*, 1014-1022. doi: 10.2307/2532087

Lukas, M., Van De Geer, S., & Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(1), 53-71. doi: 10.1111/j.1467-9868.2007.00627.x

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*(4), 661-675. doi: 10.2307/1267380

Maris, E. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523-547.

Martín, E. S., del Pino, G., & Boeck, P. D. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*, 183 - 203. doi: 10.1177/0146621605282773

McQuarrie, A. D. R., & Tsai, C.-L. (1998). *Regression and time series model selection*. World Scientific.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, *34*(3), 1436-1462. doi: 10.1214/009053606000000281

Murtaugh, P. A. (1998). Methods of variable selection in regression modeling. *Communications in Statistics - Simulation and Computation*, *27*, 711-734. doi: 10.1080/03610919808813505

Pan, J., & Shang, J. (2018). Adaptive lasso for linear mixed model selection via profile log-likelihood. *Communications in Statistics - Theory and Methods*, *47*(8), 1882-1900. doi: 10.1080/03610926.2017.1332219

Raudenbush, S., Yang, M., & Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, *9*, 141 - 157.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*, 1-25. doi: 10.18637/jss.v017.i05

Roecker, E. B. (1991). Prediction error and its estimation for subset-selected models. *Technometrics*, *33*(4), 459-468. doi: 10.2307/1403680

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461 - 464. doi: 10.1214/aos/1176344136

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, *24*, 647-656. doi: 10.1090/S0025-5718-1970-0274029-X

Shun, Z. (1997). Another look at the salamander mating data: A modified Laplace approximation approach. *Journal of the American Statistical Association*, *92*, 341-349. doi: 10.1080/01621459.1997.10473632

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion

and the finite corrections. *Communications in Statistics - Theory and Methods*, *7*(1), 13-26. doi: 10.1080/03610927808827599

Thomas, N. (1993). Asymptotic corrections for multivariate posterior moments with factored likelihood functions. *Journal of Computational and Graphical Statistics*, *2*, 309-322. doi: 10.2307/1390648

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Turing, A. (1948). Rounding-off errors in matrix processes. *Quarterly Journal of Mechanics and Applied Mathematics*, *1*, 287-308. doi: 10.1093/qjmam/1.1.287

Vaida, F., & Blanchard, S. (2005). Conditional Akaike information for mixed effects models. *Biometrika*, *92*(2), 351-370. doi: 10.1093/biomet/92.2.351

von Davier, M., & Sinharay, S. (2010). Stochastic approximation methods for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*, 174 - 193. doi: 10.3102/1076998609346970

von Davier, M., & Sinharay, S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In *In l. rutkowski (ed.), handbook of international large-scale assessment: Background, technical issues, and methods of data analysis.* (p. 155-174). Chapman Hall/CRC Press.

Wolfinger, R. D., Tobias, R., & Sall, J. P. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing*, *15*(6), 1294-1310. doi: 10.1137/0915079

Yang, Y., & Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize

learning problems. *Statistics and Computing*, *25*, 1129–1141. doi: 10.1007/s11222-014-9498-5

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*, 49–67. doi: 10.1111/j.1467-9868.2005.00532.x

Zhang, H. H., & Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, *94*(3), 691–703. doi: 10.1093/biomet/asm037

Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, *7*, 2541-2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418-1429. doi: 10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320. doi: 10.1111/j.1467-9868.2005.00503.x