Simultaneous Fixed and Random Effects Selection in Finite Mixtures of Linear Mixed-Effects Models

Ye Ting Du

Master of Science

Department of Mathematics and Statistics

McGill University

Montreal, Quebec

July 2012

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

I would like to dedicate this thesis to my mother, whose constant love, support and encouragement have always been, and will always be, my greatest source of strength and inspiration.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Professor Abbas Khalili, for proposing this interesting and challenging project, and for his guidance and support throughout the course of this work. I am particularly grateful to Professor Khalili for holding weekly meetings for the past two years, for suggesting viable avenues to solve the numerous problems I encountered, for raising constructive criticisms that were absolutely essential in improving this thesis, and for shaping me into the statistician I am today. His advice of always tackling problems from the simplest case will definitely have a profound influence on my future career.

I am especially indebted to my co-supervisor, Professor Johanna Nešlehová, who never hesitated to offer her assistance when I needed help in my research and scholarship applications, and was always there to uplift me with kind words of encouragement. Her explanation of several intricacies in the theorems greatly improved my understanding of the asymptotic properties of the proposed penalized estimators, and made the proofs much more rigorous. I would also like to take this opportunity to thank her for her collaboration on our *Metrika* paper, through which I learned a great deal about copula modeling and scientific paper writing.

I am also deeply grateful to Professor Russell Steele for providing the data from the Canadian Scleroderma Research Group registry when I was desperately seeking a real data set for this project, and for his subsequent help in interpreting the results we obtained. His encouragement is also greatly appreciated.

I would like to thank Professor David Stephens and Professor Robert Platt for generously providing the much needed additional computational resources for an extensive period of time.

System administrator Greg Lebaron's timely assistance with unicode was also especially valuable.

I am profoundly thankful to Professor David Wolfson, who convinced me to take up statistics in the first place, supervised my first research project, and coauthored my first research paper. I cannot thank him enough for his suggestions and assistance during my Ph.D. applications, and for his help in dealing with countless administrative matters, even when time was not on his side. I also greatly enjoyed the numerous insightful conversations we held on statistics, classical music, and life in general.

I also wish to thank my best friend, Vadim Roulkov, who has been a source of consistent moral support and great comfort for the past eleven years.

The warm encouragement from Yang, Angga, Maks, Phil, Tigran, Nick, Ilia and Ernest is also deeply appreciated.

I am also especially grateful to all the professors who have ever taught me.

Last but certainly not the least, I would like to thank my parents, my relatives in China, and all my friends for their love and support.

ABSTRACT

Linear mixed-effects (LME) models are frequently used for modeling longitudinal data. One complicating factor in the analysis of such data is that samples are sometimes obtained from a population with significant underlying heterogeneity, which would be hard to capture by a single LME model. Such problems may be addressed by a finite mixture of linear mixed-effects (FMLME) models, which segments the population into subpopulations and models each subpopulation by a distinct LME model. Often in the initial stage of a study, a large number of predictors are introduced. However, their associations to the response variable vary from one component to another of the FMLME model. To enhance predictability and to obtain a parsimonious model, it is of great practical interest to identify the important effects, both fixed and random, in the model. Traditional variable selection techniques such as stepwise deletion and subset selection are computationally expensive even with modest numbers of covariates and components in the mixture model. In this thesis, we introduce a penalized likelihood approach and propose a nested EM algorithm for efficient numerical computations. The estimators are shown to possess consistency and sparsity properties and asymptotic normality. We illustrate the performance of the proposed method through simulations and a real data example.

ABRÉGÉ

Les modèles linéaires mixtes (LME) sont fréquemment employés pour la modélisation des données longitudinales. Un facteur qui complique l'analyse de ce genre de données est que les échantillons sont parfois obtenus à partir d'une population d'importante hétérogénéité sous-jacente, qui serait difficile à capter par un seul LME. De tels problèmes peuvent être surmontés par un mélange fini de modèles linéaires mixtes (FMLME), qui segmente la population en sous-populations et modélise chacune de ces dernières par un LME distinct. Souvent, un grand nombre de variables explicatives sont introduites dans la phase initiale d'une étude. Cependant, leurs associations à la variable réponse varient d'un composant à l'autre du modèle FMLME. Afin d'améliorer la prévisibilité et de recueillir un modèle parcimonieux, il est d'un grand intérêt pratique d'identifier les effets importants, tant fixes qu'aléatoires, dans le modèle. Les techniques conventionnelles de sélection de variables telles que la suppression progressive et la sélection de sous-ensembles sont informatiquement chères, même lorsque le nombre de composants et de covariables est relativement modeste. La présente thèse introduit une approche basée sur la vraisemblance pénalisée et propose un algorithme EM imbriqué qui est computationnellement efficace. On démontre aussi que les estimateurs possèdent des propriétés telles que la cohérence, la parcimonie et la normalité asymptotique. On illustre la performance de la méthode proposée au moyen de simulations et d'une application sur un vrai jeu de données.

TABLE OF CONTENTS

LIST OF TABLES

Table page

CHAPTER 1
Introduction

1.1 Motivation

In longitudinal studies, data are usually collected on each subject at multiple time points in order to explore the changes of certain characteristics over time. As such, measurements taken on the same unit are generally correlated. Ignoring this correlation could lead to erroneous inference. The class of linear mixed-effects (LME) models (Laird & Ware, 1982) constitutes a powerful tool to analyze correlated data. The fixed effects of these models serve to specify the means of the observations, whereas the subject-specific random effects define the covariance structure of the observations corresponding to each individual.

In practice, data are sometimes collected from a population with significant underlying heterogeneity, i.e. each subject could belong to one of several inherent subpopulations. While LME models successfully reflect the correlation engendered from repeated measurements, they may not be able to account for the heterogeneity in the outcomes.

As a motivating example, we consider the data set on 378 patients enrolled in the Canadian Scleroderma Research Group registry with baseline visits between 2003 and 2010. Systemic sclerosis is a connective tissue disease which could cause hardening of skin, digital ulcers, Raynaud's phenomenon, shortness of breath, gastrointestinal complications and musculoskeletal problems (Wigley & Hummers, 2006). The study in question aims to identify the clinical features of systemic sclerosis (SSc) that best relate to the patients' health status and quality of life. The outcome of interest is measured by the Health Assessment Questionnaire-Disability Index (HAQ), which is calculated from a self-reported questionnaire and has been validated as an accurate measure of disability in SSc (Bruce & Fries, 2003). At each yearly clinical visit, patients complete a HAQ, which provides us with repeated measurements on each patient and therefore induces within-subject correlation. Schnitzer et al. (2011) employed an LME model to estimate the decline in HAQ over time, and concluded that the deterioration in health of individual patients might be quite heterogeneous since the standard deviation of the slopes was five times as large as the magnitude of the mean slope for the time variable. Given this substantial heterogeneity, it is of interest as to whether there was simply random variation in the decline of patients' health status, or whether we could detect a group of patients whose health status severely declined and another group whose health status only declined moderately.

A natural solution to account for both within-subject correlation and between-subject heterogeneity is to employ a finite mixture of linear mixed-effects (FMLME) models, where each component of the mixture is an LME model in itself and models one subgroup of the population. Even though such models have attracted growing attention in recent years and have been employed in many applications, the problem of variable selection in these models has received little attention.

In the initial stage of a study, a large set of covariates is usually of interest. To enhance predictability and give a parsimonious FMLME model, it is crucial to pinpoint the significant effects, both fixed and random, associated with the response variable. The omission of important fixed effects would lead to modeling bias. On the other hand, the significance of each covariate may differ from one mixture component to another, and one is prone to overfit the data by including the superfluous predictors. Care is also needed when deciding which random effects are necessary for each component. Lange and Laird (1989) and Littell, Pendergast, and Natarajan (2000) demonstrated that misspecification of covariance structure could cause substantial bias in the estimated variance of the fixed effects. Crowder (1995) also showed via examples that the use of incorrect covariance matrix could create inconsistency in parameter estimation.

Classical methods of variable selection include stepwise deletion and subset selection, coupled with a model selection criterion such as the Akaike Information Criterion or AIC (Akaike, 1972), the Bayes Information Criterion or BIC (Schwarz, 1978), and the Generalized Information Criterion or GIC (Nishii, 1984). However, these all-subset selection procedures are computationally intensive even for FMLME models with modest numbers of components and predictors. For example, the number of potential two-component FMLME submodels to be examined by these criteria is over 10^{12} for the SSc data.

The advent of more efficient techniques, such as the Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996) and the Smoothly Clipped Absolute Deviation (SCAD) by Fan and Li (2001), opens new avenues to tackle the variable selection problem. Unlike traditional methods, LASSO and SCAD achieve variable selection by estimating the effect of the non-significant covariates in the model to be exactly zero. Such methods unify parameter estimation and variable selection in one single step, and hence greatly reduce the computational burden. Khalili and Chen (2007) proposed a new regularization method for variable selection in finite mixture of regression models without random effects, and demonstrated that their method was at least as good as BIC at selecting correct models. Motivated by the success of these new regularization techniques, in this thesis, we propose a penalized likelihood approach to simultaneously identify the important fixed and random effects in an FMLME model. We also develop an efficient nested EM algorithm for numerical computations where all parameter updates in the inner M-step are of closed form, and present a procedure for tuning parameter selection. In addition, the parameter estimates are shown to possess nice asymptotic properties such as consistency, sparsity and asymptotic normality. A simulation study and the systemic sclerosis data are used to illustrate the proposed method.

The rest of the thesis is organized as follows. In the remainder of this chapter, we provide a brief review of the following key concepts involved in the methodology that we will introduce later on: linear mixed-effects models, finite mixture models, EM algorithm, and model selection. In Chapter 2, we define FMLME models formally. Chapter 3 provides a review of existing feature selection methods in models with random effects, and proposes a penalized likelihood approach to simultaneously select fixed and random effects in FMLME models. In Chapter 4, we present our numerical algorithm and a data adaptive method for choosing tuning parameters. Chapter 5 discusses the large-sample properties of the penalized maximum likelihood estimators. Chapter 6 presents a simulation study to investigate the finite sample performance of our method. In Chapter 7, we apply the proposed method to the systemic sclerosis data set. Chapter 8 contains conclusions and suggestions for future research.

1.2 Linear mixed-effects models

We begin by introducing the class of linear mixed-effects (LME) models, which are frequently employed to model longitudinal data. Suppose we have a set of observations on m subjects, with n_i repeated measurements $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})'$ on the ith subject. Let $\boldsymbol{Y}_i$ denote the random vector corresponding to $\boldsymbol{y}_i$. The LME model is characterized by a combination of

- $n_i \times p$ fixed effects design matrix $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})'$, with $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})'$
- $n_i \times q$ random effects design matrix $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i})'$, with $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijq})'$
- $p \times 1$ vector of population-specific fixed effects $\boldsymbol{\beta}$ (i.e. the same for all subjects)
- $q \times 1$ vector of subject-specific random effects $\boldsymbol{\alpha}_i$
- $n_i \times 1$ vector of errors $\boldsymbol{\varepsilon}_i$

The vector representation of the LME model is given by

$$\begin{cases} \boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\alpha}_i \sim \mathcal{N}(0, \sigma^2\boldsymbol{\Psi}) \\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_{n_i}) \\ \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m \text{ independent} \end{cases} . \tag{1.1}$$

From (1.1), it is easy to see that $\mathrm{E}\left[\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{Z}_i\right]=\boldsymbol{X}_i\boldsymbol{\beta}$ and $\mathrm{Var}\left[\boldsymbol{Y}_i|\boldsymbol{X}_i,\boldsymbol{Z}_i\right]=\sigma^2(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i'+\boldsymbol{I}_{n_i})$. It is precisely by modeling the covariance matrix $\sigma^2(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i'+\boldsymbol{I}_{n_i})$ that we account for the correlation between observations from the same subject. We also notice that $\mathrm{Cov}\left[\boldsymbol{Y}_i,\boldsymbol{Y}_j\right]=0$ if $i\neq j$, which implies that observations are independent between different subjects.

In order to carry out likelihood-based inference about the unknown parameters $(\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)$, we find the marginal density function of $\boldsymbol{Y}_i$ by integrating out the random effects $\boldsymbol{\alpha}_i$, namely,

$$f(\boldsymbol{y}_i;\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)=\int f(\boldsymbol{y}_i|\boldsymbol{\alpha}_i;\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)f(\boldsymbol{\alpha}_i;\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)\,\mathrm{d}\boldsymbol{\alpha}_i.$$

Note that both density functions inside the integral are multivariate normal. Since the convolution of normal densities is normal, the marginal density function of $\boldsymbol{Y}_i$ is given by the multivariate normal

$$f(\boldsymbol{y}_i;\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)=\mathcal{N}\big(\boldsymbol{y}_i;\boldsymbol{X}_i\boldsymbol{\beta},\sigma^2(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i'+\boldsymbol{I}_{n_i})\big). \tag{1.2}$$

The log-likelihood formed from repeated measurements on m subjects is thus given by

$$l_m(\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)=\sum_{i=1}^{m}\left\{-\frac{1}{2}\log\left|\sigma^2(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i'+\boldsymbol{I}_{n_i})\right|-\frac{1}{2\sigma^2}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})'(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i'+\boldsymbol{I}_{n_i})^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})\right\}.$$

One can then proceed to find the maximum likelihood (ML) estimates of $(\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)$ by maximizing l_m with respect to these parameters. The ML estimators have a number of desirable properties, most notably consistency and asymptotic normality. However, one drawback of the ML estimators is that they tend to slightly underestimate the variance components. This problem can be corrected for by using the following restricted maximum likelihood (REML) derived by Harville (1974)

$$rl_m(\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)=\sum_{i=1}^{m}\left\{-\frac{1}{2}\log\left|\frac{1}{\sigma^2}\boldsymbol{X}_i'(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i'+\boldsymbol{I}_{n_i})^{-1}\boldsymbol{X}_i\right|\right\}+l_m(\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2).$$

Despite the advantage of REML estimators, we elect to use the ML estimators in the methodology we develop later on, since the extra term in $rl_m(\boldsymbol{\beta},\boldsymbol{\Psi},\sigma^2)$ would present a significant obstacle in the estimation procedure. This complication outweighs the benefit of using REML estimators, because the difference between ML and REML estimates is immaterial when the sample size is large, where a rule of thumb for "large" is $m > 30$, as suggested by Snijders and Bosker (2003). Furthermore, in the context of LME models, Gurka (2006) presented an extensive simulation study on model selection using criteria such as AIC and BIC, and demonstrated that the performance of ML and REML estimators are quite similar.

Popular numerical algorithms for maximizing $l_m(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2)$ and $rl_m(\boldsymbol{\beta}, \boldsymbol{\Psi}, \sigma^2)$ include Newton-Raphson method and the expectation-maximization (EM) algorithm of Dempster, Laird, and Rubin (1977). In the context of LME models, Lindstrom and Bates (1988) gave a thorough description of the Newton-Raphson algorithm along with the necessary derivatives, and Laird, Lange, and Stram (1987) provided a detailed computational procedure using the EM algorithm.

For a comprehensive treatment of the theory behind LME models and their applications, one could turn to the book by Verbeke and Molenberghs (2009).

1.3 Finite mixture models

The LME model (1.1) assumes that given the covariate values, the same mean expression and variance structure apply to all subjects from the population of interest. In practice, however, data may be collected from a population with substantial underlying heterogeneity. Namely, the overall population could comprise several inherent subpopulations. Therefore, one universal model would be insufficient to describe the random behavior of observations from this kind of population. Such problems may be tackled by employing finite mixture models, which have been extensively studied in the literature. See, for example, McLachlan and Peel (2000).

For simplicity, we assume in this section that the finite mixture model density function does not depend on any covariates. Consider a population composed of K subpopulations.

Suppose we have observed data $(\boldsymbol{y}_1', \dots, \boldsymbol{y}_m')'$ from m subjects, and it is unknown which subpopulation each $\boldsymbol{y}_i$ belongs to. The corresponding K-component parametric finite mixture model states that the density function of each $\boldsymbol{y}_i$ is given by the following convex combination of K probability density functions

$$f(\boldsymbol{y}_i; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k), \quad \text{where } \sum_{k=1}^{K} \pi_k = 1. \tag{1.3}$$

In (1.3), the mixing proportions π_k's represent the contribution of each component-wise density $f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k)$ to the overall density $f(\boldsymbol{y}_i; \boldsymbol{\Phi})$, and $\boldsymbol{\Phi} = (\pi_1, \dots, \pi_{K-1}, \boldsymbol{\phi}_1', \dots, \boldsymbol{\phi}_K')'$ contains all the unknown parameters in the mixture model. Also, the conditional distribution function of $\boldsymbol{Y}_i$ given membership of the k-th subpopulation is $f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k)$.

The log-likelihood formed from the entire observational vector $(\boldsymbol{y}_1', \dots, \boldsymbol{y}_m')'$ is given by

$$\begin{aligned} \log L(\boldsymbol{\Phi}) &= \sum_{i=1}^{m} \log f(\boldsymbol{y}_i; \boldsymbol{\Phi}) \\ &= \sum_{i=1}^{m} \log \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k). \end{aligned} \tag{1.4}$$

The ML estimate of $\boldsymbol{\Phi}$ can then be found by maximizing log-likelihood (1.4).

In some applications, one may be able to find the ML estimates of the parameters of interest by solving the score function directly. However, this is not the case for finite mixture models. The log-likelihood (1.4) involves the log of a sum, which makes the derivative computationally intractable. The EM algorithm provides a much more convenient method for finding ML estimates of parameters in finite mixture models, and will be recapitulated next.

1.4 EM algorithm

When maximizing the log-likelihood of a statistical model, classical optimization methods such as Newton-Raphson, Quasi-Newton and conjugate gradient usually work with the incomplete-data log-likelihood, namely the log-likelihood formed from the observed data only. The EM algorithm, on the other hand, is an iterative method that maximizes the likelihood by working with the complete-data likelihood, which involves both observed and unobserved data.

In brief, each iteration of the EM algorithm consists of an E-step which computes the conditional expectation of the log-likelihood given the current parameter estimates, and an M-step which updates the parameters by maximizing this expected log-likelihood. We alternate between the E- and M-steps until convergence is reached. The theory of EM algorithm guarantees that the observed likelihood is increased at each iteration (see Wu, 1983). Therefore the parameter estimates must converge to a stationary point, provided that the likelihood is bounded.

We now describe in more detail the EM algorithm in the context of finite mixture models, since this will form part of the nested EM algorithm that we develop later on. Suppose we have an observed data vector $(\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')'$, then the corresponding unobserved data would be the associated component label vectors $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_m$, where each $\boldsymbol{t}_i$ is a K-dimensional vector whose k-th entry, t_{ik}, is equal to 1 if $\boldsymbol{y}_i$ arose from the k-th component of the mixture or 0 otherwise. The complete-data log-likelihood for $\boldsymbol{\Phi}$ is given by

$$l^c(\boldsymbol{\Phi}) = \sum_{k=1}^{K} \sum_{i=1}^{m} t_{ik} \{\log \pi_k + \log f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k)\}. \tag{1.5}$$

Let ν index the EM iterations.

In the <u>E-step</u>:

we compute the conditional expectation of the complete-data log-likelihood, which is given by

$$\boldsymbol{Q}(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(\nu)}) = \left\{ \sum_{k=1}^{K} \sum_{i=1}^{m} \tau_{ik}^{(\nu)} \log \pi_k \right\} + \left\{ \sum_{k=1}^{K} \sum_{i=1}^{m} \tau_{ik}^{(\nu)} \log f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k) \right\}, \tag{1.6}$$

where $\tau_{ij}^{(\nu)}$ is the posterior probability that observation $\boldsymbol{y}_i$ belongs to the j-th component given the current parameter estimates, namely

$$\tau_{ij}^{(\nu)} = \frac{\pi_j^{(\nu)} f_j(\boldsymbol{y}_i; \boldsymbol{\phi}_j^{(\nu)})}{\sum_{k=1}^{K} \pi_k^{(\nu)} f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k^{(\nu)})}.$$

This concludes the E-step.

In the M-step:

we maximize equation (1.6) with respect to $\boldsymbol{\Phi}$. It could be shown that the updates of π_k's are given in closed form by

$$\pi_k^{(\nu+1)} = \sum_{i=1}^{m} \tau_{ik}^{(\nu)}/m, \qquad (k = 1, \ldots, K). \tag{1.7}$$

To update the $\boldsymbol{\phi}_k$'s, we have to solve the system of K equations

$$\sum_{k=1}^{K} \sum_{i=1}^{m} \tau_{ik}^{(\nu)} \frac{\partial}{\partial \boldsymbol{\phi}_k} \log f_k(\boldsymbol{y}_i; \boldsymbol{\phi}_k) = 0, \qquad (k = 1, \ldots, K).$$

This concludes the M-step.

For a more detailed derivation of the above results, one could consult the book by McLachlan and Krishnan (1997) and references therein.

Note that the EM algorithm can also be used to perform parameter estimation in the LME model (1.1). In this case, the unobserved data would be the random effects $\boldsymbol{\alpha}_i$'s. We shall also use this idea later on in our nested EM algorithm.

1.5 Variable selection in regression models

It is often of interest to identify the most important variables in statistical modeling. In the context of regression analysis, this is termed variable selection or feature selection. It is desirable to produce parsimonious models, because they usually offer greater interpretability than models with high complexity. Furthermore, parsimonious models also possess enhanced predictive power. It is a well known fact that one can increase the likelihood and reduce the prediction bias by including more explanatory variables in the model, but this would also make the prediction variance larger (see, for example, Seber & Lee, 2003).

Many researchers in the past have investigated model selection techniques in linear regression and generalized linear models, so as to find the right balance in the bias-variance tradeoff. Notable classical methods include the AIC and BIC. Given the regression parameter estimates $\hat{\boldsymbol{\beta}}$ of a particular model, these two criteria are given by

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + 2 \cdot df \tag{1.8}$$

$$\text{and} \qquad BIC = -2l(\hat{\boldsymbol{\beta}}) + \log(N) \cdot df, \tag{1.9}$$

where $l(\hat{\boldsymbol{\beta}})$ is the estimated log-likelihood, and the degree of freedom df is the number of elements in $\boldsymbol{\beta}$. In both criteria, the first term measures the model goodness, whereas the second term consists of a penalty on model complexity.

In order to choose an optimal model, one would need to compute the AIC or BIC for all possible sub-models, which could be computationally taxing. For example, if p potentially significant regressors were available, the number of candidate models to be examined by AIC or BIC would be 2^p.

An alternative to subset selection methods is to use regularization techniques such as LASSO (Tibshirani, 1996), SCAD (Fan & Li, 2001), and Adaptive LASSO (Zou, 2006). Unlike AIC and BIC which apply penalties on the number of parameters, these new techniques apply penalties on the parameters themselves, and estimate the effect of the non-significant covariates to be exactly zero. Parameter estimation and variable selection are thus achieved simultaneously, which significantly reduces the computational cost. This motivated us to employ these techniques in our applications which will be presented in later chapters. We now define these penalty functions formally. Let λ denote the tuning parameter. For a given generic parameter β in the model, the three penalty functions are given by

(a) LASSO penalty:

$$p(\beta) = \lambda \cdot |\beta|.$$

(b) Adaptive LASSO penalty:

$$p(\beta) = \lambda \cdot w|\beta|,$$

for some adaptive weight w.

(c) SCAD penalty:

$$p'(\beta) = \lambda_{mk} \left\{ \mathbf{1}(|\beta| \leq \lambda_{mk}) + \frac{(a\lambda_{mk} - |\beta|)_+}{(a-1)\lambda_{mk}} \mathbf{1}(|\beta| > \lambda_{mk}) \right\}, \qquad \text{for some } a > 2.$$

For the SCAD penalty, although the pair (λ, a) could be chosen over a two-dimensional grid using criteria such as cross-validation, this procedure may be computationally intensive. Fan and Li (2001) showed using a Bayes risks argument that $a = 3.7$ is a good choice for various variable selection problems in practice.

All these three penalty functions are equal to zero at $\beta = 0$, and non-differentiable at the origin. It is precisely this feature that allows them to reduce small estimated coefficients to zero and achieve sparsity (consistent variable selection). Despite sharing these common features, these penalties do behave slightly differently. For example, the LASSO penalty tends to shrink all effects by similar amounts. In this regard, the Adaptive LASSO penalty could be considered as a refinement of the LASSO penalty, since by introducing the adaptive weights w, it can apply a heavy shrinkage to the zero parameters and leave the significant parameters relatively unpenalized, which leads to improved parameter estimation and selection properties. In particular, Zou (2006) showed that if the adaptive weight w is chosen to be the inverse of a consistent estimator of β, then root-n consistency and sparsity can be simultaneously achieved for suitable choices of tuning parameter. In this thesis, we will use the inverse of the ML estimates as the adaptive weights.

CHAPTER 2
Finite mixture of linear mixed-effects (FMLME) models

In the previous chapter, we discussed that LME models are designed to take into consideration the correlation in repeated measurements, and that finite mixture models can successfully account for the heterogeneity in the observations. Therefore, when confronted with data with both correlation and substantial underlying heterogeneity, it seems natural to combine these two modeling ideas by creating a finite mixture of linear mixed-effects (FMLME) models, which segments the overall population into several subpopulations, and each of these subpopulations would call for its own LME modeling between the response and explanatory variables. We shall now formally define FMLME models.

2.1 Model specification

Suppose we have m subjects, with n_i, $i = 1, \ldots, m$, measurements from each subject. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ denote the $n_i \times 1$ response vector for subject i, and $\boldsymbol{Y}_i$ the corresponding random vector. Let $N = \sum_{i=1}^{m} n_i$ be the total number of observations. Let $\boldsymbol{X}_i$ be the corresponding $n_i \times p$ design matrix of fixed effects, and $\boldsymbol{Z}_i$ the $n_i \times q$ design matrix of random effects. We assume that observations from different individuals, namely the $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$'s, are independent.

Definition 1. In a heterogenous population consisting of K subpopulations, we say that $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ follows a *finite mixture of linear mixed-effects models* of order K if the marginal density function of $\boldsymbol{Y}_i$, given $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, has the form

$$f(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{y}_i; \boldsymbol{X}_i \boldsymbol{\beta}_k, \sigma_k^2 (\boldsymbol{Z}_i \boldsymbol{\Psi}_k \boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})\right), \tag{2.1}$$

where $\mathbf{\Phi}$ contains all the parameters in the mixture density, $\mathcal{N}(.;\mu,\Sigma)$ denotes the multivariate normal distribution with mean vector μ and variance-covariance matrix Σ, $\pi_k > 0$ for all $k = 1,\ldots,K$, and $\sum_{k=1}^{K} \pi_k = 1$.

In the special case of $K = 1$, we obtain the marginal distribution (1.2) of $\boldsymbol{Y}_i$ in the usual LME model. When $K > 1$, the parameters $\boldsymbol{\beta}_k$, $\mathbf{\Psi}_k$ and σ_k^2 are assumed to be different from one component to another, therefore each mixture component has a distinct mean and variance.

FMLME models have attracted considerable attention in recent years. Verbeke and Lesaffre (1996) investigated the impact of normality assumptions for the random effects on their estimates in the LME model when the random effects are actually distributed according to a mixture of normal densities. Yau, Lee, and Ng (2003) applied a mixture of two LME models to analyze neonatal hospital length of stay, where the two components correspond to the short-stay and the long-stay patients, respectively. Celeux, Martin, and Lavergne (2005) and Ng et al. (2006) utilized mixture of LME models to perform clustering of correlated gene expression profiles. Scharl, Grün, and Leisch (2010) evaluated the differences between mixtures of regression models with and without random effects. Martella et al. (2011) performed classification of sibling pairs using finite mixture models with random effects.

Given the finite mixture structure of FMLME models, any meaningful analysis should be predicated on the assumption that the model is identifiable.

Definition 2. Consider a finite mixture of linear mixed-effects model of the form (2.1). For given design matrices $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m)$ and $(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_m)$, the finite mixture of linear mixed-effects model is said to be *identifiable* if for any two parameter vectors $\mathbf{\Phi}$ and $\mathbf{\Phi}^*$,

$$\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k, \sigma_k^2(\boldsymbol{Z}_i\mathbf{\Psi}_k\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})\right) = \sum_{k=1}^{K^*} \pi_k^* \mathcal{N}\left(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k^*, \sigma_k^{2*}(\boldsymbol{Z}_i\mathbf{\Psi}_k^*\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})\right)$$

for each $i = 1,\ldots,m$ and all possible values of $\boldsymbol{y}_i$, implies that $K = K^*$ and $\mathbf{\Phi} = \mathbf{\Phi}^*$ up to a permutation.

The identifiability of finite mixtures of regression models depends on the following three factors: the component-wise density, the number of components K, and the design matrices. For finite mixtures of linear regression models with fixed designs, Hennig (2000) demonstrated that a sufficient condition for identifiability is that the number of components K is less than the minimum number of $(h_x - 1)$-dimensional hyperplanes formed by the design points, where h_x is the number of predictors in $\boldsymbol{X}$ excluding the intercept. Let $\boldsymbol{\Upsilon}$ be the matrix formed by the distinct columns of the following matrix

$$\begin{bmatrix} \boldsymbol{X}_1, \boldsymbol{Z}_1 \\ \boldsymbol{X}_2, \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{X}_m, \boldsymbol{Z}_m \end{bmatrix}.$$

Then in our context, Hennig's condition translates to the restriction that K must be smaller than the number of $(h_{\Upsilon} - 1)$-dimensional hyperplanes covered by the rows of $\boldsymbol{\Upsilon}$. In the special case where the $\boldsymbol{Z}_i$'s are a subset of the $\boldsymbol{X}_i$'s, this condition becomes a condition on the design points of the $\boldsymbol{X}_i$'s only. Heuristically, this condition means that if the design points exhibit too little variability from one subject to another, for example when all the variables are categorical indicators, then identifiability problems could ensue. In the subsequent theoretical exposition, we assume that the FMLME model in question is identifiable. In the simulation study and real data example, our model can be verified to be identifiable.

The next chapters are devoted to simultaneous fixed and random effects selection in FMLME models, which is the main topic of this thesis.

CHAPTER 3
Simultaneous fixed and random effects selection in FMLME models

Variable selection in models with random effects has gathered considerable attention in the past. One major difficulty in such problems is precisely the selection of random effects. Traditionally, this is done by performing hypothesis tests (Wald test or likelihood ratio test) on nested models with different covariance structures. However, these tests suffer from the fact that the null hypotheses of interest are on the boundary of the parameter space, which violates the regularity conditions and thus renders the asymptotic results of these tests unapplicable (Verbeke & Molenberghs, 2009). This provided an impetus for researchers to investigate alternative methods to perform variable selection in such models. Chen and Dunson (2003) proposed a Bayesian approach to select the random effects in an LME model by assuming the fixed effects component of the model is known. Pu and Niu (2006) employed an extended GIC criterion to select first the fixed effects by keeping all the random effects in the model, and then select the random effects by keeping the chosen fixed effects from the previous step. Recently, Bondell, Krishna, and Ghosh (2010) proposed to simultaneously select fixed and random components in an LME model via penalized likelihood, and demonstrated that simultaneous selection provides better performance than choosing the fixed and random components sequentially. Ibrahim et al. (2011) also proposed a penalized likelihood approach to simultaneously select fixed and random effects in generalized linear mixed models. Their estimation procedure, however, involves Markov chain Monte Carlo simulation, which could be time consuming.

Motivated by the work of Chen and Dunson (2003) and Bondell et al. (2010), we propose a new method for joint selection of fixed and random effects in FMLME models. By applying the modified Cholesky decomposition proposed by these authors, we can factorize the covariance matrix of the random effects from the k-th component of the FMLME model (2.1) as $\boldsymbol{\Psi}_k = \boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'\boldsymbol{D}_k$, where $\boldsymbol{D}_k = \mathrm{diag}(d_{k1}, d_{k2}, \ldots, d_{kq})$ is a diagonal matrix, and $\boldsymbol{\Gamma}_k$ is a $q \times q$ lower triangular matrix with 1's on the diagonal. Given this decomposition, conditioning on membership of the k-th mixture component, the response vector of the i-th subject is governed by the component-wise LME model

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta}_k + \boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{b}_{ki} + \boldsymbol{\varepsilon}_{ki},$$

where $\boldsymbol{b}_{ki} = (b_{ki_1}, \ldots, b_{ki_q})'$ is the new $q \times 1$ random effect vector with distribution $\mathcal{N}(0, \sigma_k^2\boldsymbol{I}_q)$, and $\boldsymbol{\varepsilon}_{ki}$ is an $n_i \times 1$ error vector with distribution $\mathcal{N}(0, \sigma_k^2\boldsymbol{I}_{n_i})$. The advantage of this decomposition is that, if d_{kl} is estimated to be 0, then the l-th row and column of the resulting covariance matrix $\boldsymbol{\Psi}_k$ are 0, which facilitates the selection of random effects. Furthermore, there is no hypothesis testing involved in this selection procedure, so we effectively circumvent the complications discussed at the beginning of this chapter. Note that the parameters in $\boldsymbol{D}_k$ and $\boldsymbol{\Gamma}_k$ are functionally related, in that once d_{kl} is identified to be 0, all the parameters in the l-th row of $\boldsymbol{\Gamma}_k$ have to be set to 0. This ensures the identifiability of the parameters in $\boldsymbol{\Gamma}_k$ in the estimation process.

Applying the above variance decomposition to each component of the FMLME model, we can reformulate the mixture density (2.1) as

$$\begin{aligned} f(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\Phi}) &= \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{y}_i; \boldsymbol{X}_i\boldsymbol{\beta}_k, \sigma_k^2(\boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'\boldsymbol{D}_k\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})\right) \\ &= \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_k), \end{aligned} \tag{3.1}$$

where $\boldsymbol{\Phi} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_K', \boldsymbol{d}_1', \ldots, \boldsymbol{d}_K', \gamma_1', \ldots, \boldsymbol{\gamma}_K', \sigma_1^2, \ldots, \sigma_K^2)'$ denotes the vector of all unknown parameters for the mixture density f, and $\boldsymbol{\phi}_k = (\boldsymbol{\beta}_k', \boldsymbol{d}_k', \boldsymbol{\gamma}_k', \sigma_k^2)'$ denotes the vector of parameters for the k-th component density of the mixture, f_k. In the k-th component variance, $\boldsymbol{D}_k = \text{diag}(\boldsymbol{d}_k') = \text{diag}(d_{k1}, \dots, d_{kq})$, and $\boldsymbol{\Gamma}_k$ is a $q \times q$ lower triangular matrix whose off-diagonal elements are given by the vector $\boldsymbol{\gamma}_k$.

Let $(\boldsymbol{y}_1, \boldsymbol{X}_1, \boldsymbol{Z}_1), (\boldsymbol{y}_2, \boldsymbol{X}_2, \boldsymbol{Z}_2), \dots, (\boldsymbol{y}_m, \boldsymbol{X}_m, \boldsymbol{Z}_m)$ be a random sample of observations governed by the FMLME model (3.1). The log-likelihood for $\boldsymbol{\Phi}$ formed from the observed sample is given by

$$l_m(\boldsymbol{\Phi}) = \sum_{i=1}^{m} \log\{\sum_{k=1}^{K} \pi_k f_k(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_k)\}. \tag{3.2}$$

To perform fixed and random effects selection, we aim to maximize the following penalized log-likelihood function

$$pl_m(\boldsymbol{\Phi}) = l_m(\boldsymbol{\Phi}) - \boldsymbol{p}_m(\boldsymbol{\Phi}), \tag{3.3}$$

with the penalty

$$\boldsymbol{p}_m(\boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k \cdot m \left(\sum_{j=1}^{p} p_{mk}(\beta_{kj}) + \sum_{j=1}^{q} p_{mk}(d_{kj}) \right), \tag{3.4}$$

where the $p_{mk}(\theta)$ are non-negative and non-decreasing functions in $|\theta|$. By maximizing the log-likelihood with a penalty on the β_{kj}'s and d_{kj}'s, the coefficient of some fixed effects and the variance of some random effects would be estimated to be zero. Simultaneous exclusion of these effects from the model is then achieved. The penalties imposed on the parameters from the k-th mixture component are chosen to be proportional to $\pi_k \cdot m$, the virtual sample size of the k-th subpopulation.

In the present thesis, we will consider the LASSO, Adaptive LASSO and SCAD penalties introduced in Section 1.5 of Chapter 1. Let λ_{mk} denote the tuning parameter. The subscript m emphasizes that the tuning parameter is dependent on the sample size, and the subscript k indicates that each mixture component has a different tuning parameter. In the context of FMLME models, these three penalty functions are defined as follows,

(a) LASSO penalty:

$$p_{mk}(\theta) = \lambda_{mk}|\theta|.$$

(b) Adaptive LASSO penalty:

$$p_{mk}(\theta) = \lambda_{mk} \cdot w|\theta|,$$

where we choose the adaptive weights w to be the inverse of ML estimates.

(c) SCAD penalty:

$$p'_{mk}(\theta) = \lambda_{mk} \left\{ \mathbf{1}(|\theta| \leq \lambda_{mk}) + \frac{(a\lambda_{mk} - |\theta|)_+}{(a-1)\lambda_{mk}} \mathbf{1}(|\theta| > \lambda_{mk}) \right\}, \qquad \text{for some } a > 2,$$

where we follow Fan and Li (2001) and use $a = 3.7$ for our applications.

For more discussions on these three penalty functions, we refer the reader to Section 1.5 of Chapter 1.

In the next chapter, we present an efficient numerical algorithm to perform the maximization of the penalized log-likelihood (3.3).

CHAPTER 4
Numerical algorithm for variable selection and estimation

The EM algorithm introduced in Section 1.4 has been a popular method for parameter estimation in the context of finite mixtures, when the number of components K is given. We propose a nested EM algorithm to maximize the penalized log-likelihood (3.3). Concisely, the outer E-step is due to the mixture structure, and computes the posterior probability of each observation belonging to one component of the mixture. The outer M-step is an EM algorithm in itself (inner EM), which maximizes the conditional expectation of the log-likelihood for each component-wise LME model. We now describe our nested EM algorithm in more detail.

Outer E-step:

In order to apply the EM algorithm, we first need to formulate our problem as an incomplete-data problem: consider $\{(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) : i = 1, \dots, m\}$ as the observed data and their associated component-label vectors $\boldsymbol{t}_1, \dots, \boldsymbol{t}_m$ as unobserved data, where each $\boldsymbol{t}_i$ is a K-dimensional vector whose k-th entry, t_{ik}, is equal to 1 if $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ arose from the k-th component of the mixture or 0 otherwise. The penalized complete-data log-likelihood for $\boldsymbol{\Phi}$ is given by

$$pl^c(\boldsymbol{\Phi}) = \sum_{k=1}^{K} \sum_{i=1}^{m} t_{ik} \{\log \pi_k + \log f_k(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_k)\} - \boldsymbol{p}_m(\boldsymbol{\Phi}). \tag{4.1}$$

Upon taking the conditional expectation of (4.1), we obtain the objective function for the outer M-step,

$$\boldsymbol{Q}(\boldsymbol{\Phi}; \boldsymbol{\Phi}^{(\nu)}) = \left\{ \sum_{k=1}^{K} \sum_{i=1}^{m} \tau_{ik}^{(\nu)} \log \pi_k \right\} + \left\{ \sum_{k=1}^{K} \sum_{i=1}^{m} \tau_{ik}^{(\nu)} \log f_k(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_k) - \boldsymbol{p}_m(\boldsymbol{\Phi}) \right\}, \tag{4.2}$$

where ν indexes the outer EM iterations, and $\tau_{ij}^{(\nu)}$ is given by

$$\tau_{ij}^{(\nu)} = \frac{\pi_j^{(\nu)} f_j\left(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_j^{(\nu)}\right)}{\sum_{k=1}^{K} \pi_k^{(\nu)} f_k\left(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_k^{(\nu)}\right)}.$$

Outer M-step:

The M-step of the outer EM algorithm consists of updating current parameter estimates by maximizing the objective function (4.2) with respect to $\boldsymbol{\Phi}$. In the classical EM algorithm presented in Section 1.4, the mixing proportions are updated by

$$\pi_k^{(\nu+1)} = \sum_{i=1}^{m} \tau_{ik}^{(\nu)}/m, \qquad (k = 1, \ldots, K). \tag{4.3}$$

Note that this only maximizes the leading term in (4.2). Maximizing (4.2) itself with respect to π_k would be more complex. For simplicity, we update π_k according to (4.3), which worked well in our simulations. One justification for this is that we could simply replace the π_k involved in $\boldsymbol{p}_m(\boldsymbol{\Phi})$ by its estimate from the previous iteration, namely $\pi_k^{(\nu)}$, therefore the entire second term in (4.2) could be regarded as a constant when taking derivative with respect to π_k.

To update the other parameters in $\boldsymbol{\Phi}$, namely $\{\boldsymbol{\phi}_k : k = 1, \ldots K\}$, we need to maximize the second term in (4.2), viz.

$$\arg\max_{\boldsymbol{\Phi}} \sum_{k=1}^{K} \sum_{i=1}^{m} \tau_{ik}^{(\nu)} \log f_k(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_k) - \boldsymbol{p}_m(\boldsymbol{\Phi}). \tag{4.4}$$

However, since the parameters $\boldsymbol{\phi}_k$ are mathematically independent from one component to another, this is equivalent to the component-wise maximization

$$\boldsymbol{\phi}_k^{(\nu+1)} = \arg\max_{\boldsymbol{\phi}_k} \sum_{i=1}^{m} \tau_{ik}^{(\nu)} \log f_k(\boldsymbol{y}_i; \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi}_k) - \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k), \qquad (k = 1, \ldots, K), \tag{4.5}$$

where $\boldsymbol{p}_{mk}(\boldsymbol{\phi}_k) = \pi_k^{(\nu+1)} \cdot m \left\{ \sum_{j=1}^{p} p_{mk}(\beta_{kj}) + \sum_{j=1}^{q} p_{mk}(d_{kj}) \right\}$. Again, we use an EM algorithm to carry out the maximization of (4.5). This is the "inner EM", which we describe next.

Complete likelihood for Inner EM:

Suppose we are performing maximization for the k-th component. We first translate this into an incomplete-data problem by treating the random effects $\boldsymbol{b}_k = (\boldsymbol{b}'_{k1}, \ldots, \boldsymbol{b}'_{km})'_{mq\times 1}$ as unobserved, where the subscript k indicates the component. This notation is necessary since for each of the mixture components, the random effects $\boldsymbol{b}_k$ are assumed to have a different distribution.

For simplicity of notation, in the following exposition, we do not annotate explicitly the dependence of the various conditional distribution functions on the design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, which are understood to be given and fixed.

We next find the complete-data likelihood for the i-th subject $L_k^c(\boldsymbol{\phi}_k|\boldsymbol{y}_i, \boldsymbol{b}_{ki})$,

$$\begin{aligned}
L_k^c(\boldsymbol{\phi}_k|\boldsymbol{y}_i, \boldsymbol{b}_{ki}) &= f(\boldsymbol{y}_i|\boldsymbol{b}_{ki}, \boldsymbol{\phi}_k) \cdot f(\boldsymbol{b}_{ki}|\boldsymbol{\phi}_k) \\
&= \frac{1}{(2\pi\sigma_k^2)^{n_i/2}} \exp\left(-\frac{1}{2\sigma_k^2} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k - \boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{b}_{ki})'\boldsymbol{I}_{n_i}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k - \boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{b}_{ki}) \right) \\
&\quad \times \frac{1}{(2\pi\sigma_k^2)^{q/2}} \exp\left(-\frac{1}{2\sigma_k^2} \boldsymbol{b}'_{ki}\boldsymbol{I}_q\boldsymbol{b}_{ki} \right) \\
&= \frac{1}{(2\pi\sigma_k^2)^{\frac{n_i+q}{2}}} \exp\left(-\frac{1}{2\sigma_k^2} ||\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k - \boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{b}_{ki}||^2 - \frac{1}{2\sigma_k^2} \boldsymbol{b}'_{ki}\boldsymbol{b}_{ki} \right). \qquad (4.6)
\end{aligned}$$

Therefore, the complete-data log-likelihood for the i-th subject is found to be

$$l_k^c(\boldsymbol{\phi}_k|\boldsymbol{y}_i, \boldsymbol{b}_{ki}) = -\frac{n_i+q}{2} \log \sigma_k^2 - \frac{1}{2\sigma_k^2} \Big(||\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k - \boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{b}_{ki}||^2 + \boldsymbol{b}'_{ki}\boldsymbol{b}_{ki} \Big).$$

Given the above complete-data log-likelihood for a single subject, by incorporating the weights $\tau_{ik}^{(\nu)}$ and adding the penalty, then summing over i, we obtain the penalized complete-data log-likelihood for the k-th component,

$$pl_k^c(\boldsymbol{\phi}_k) = \sum_{i=1}^{m} \tau_{ik}^{(\nu)} l_k^c(\boldsymbol{\phi}_k|\boldsymbol{y}_i, \boldsymbol{b}_{ki}) - \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k). \qquad (4.7)$$

Equation (4.7) is similar to equation (4.5), except we replaced the observed-data log-likelihood by the corresponding complete-data log-likelihood. Instead of directly maximizing (4.5), we iterate between finding the conditional expectation of $\boldsymbol{b}_k$ and maximizing the conditional expectation of (4.7) until convergence. The theory of EM algorithm guarantees that at least a local maximum of (4.5) would be found in this fashion. Expanding equation (4.7), we get

$$\begin{aligned} pl_k^c(\boldsymbol{\phi}_k) = \sum_{i=1}^m & \left\{ -\tau_{ik}^{(\nu)} \frac{n_i+q}{2} \log \sigma_k^2 - \frac{\tau_{ik}^{(\nu)}}{2\sigma_k^2} \left(||\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_k - \boldsymbol{Z}_i \boldsymbol{D}_k \boldsymbol{\Gamma}_k \boldsymbol{b}_{ki}||^2 + \boldsymbol{b}_{ki}' \boldsymbol{b}_{ki} \right) \right\} \\ & - \pi_k \cdot m \left(\sum_{j=1}^p p_{mk}(\beta_{kj}) + \sum_{j=1}^q p_{mk}(d_{kj}) \right). \end{aligned} \tag{4.8}$$

Dropping out the terms that do not involve either $\boldsymbol{\beta}_k$ or $\boldsymbol{d}_k$ in (4.8), this is then equivalent to *minimizing* the conditional expectation of

$$\begin{aligned} & \sum_{i=1}^m \tau_{ik}^{(\nu)} ||\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_k - \boldsymbol{Z}_i \boldsymbol{D}_k \boldsymbol{\Gamma}_k \boldsymbol{b}_{ki}||^2 + 2\sigma_k^2 \cdot \pi_k \cdot m \left(\sum_{j=1}^p p_{mk}(\beta_{kj}) + \sum_{j=1}^q p_{mk}(d_{kj}) \right) \\ = & \sum_{i=1}^m \left|\left| \sqrt{\tau_{ik}^{(\nu)}} \boldsymbol{y}_i - \sqrt{\tau_{ik}^{(\nu)}} \boldsymbol{X}_i \boldsymbol{\beta}_k - \sqrt{\tau_{ik}^{(\nu)}} \boldsymbol{Z}_i \boldsymbol{D}_k \boldsymbol{\Gamma}_k \boldsymbol{b}_{ki} \right|\right|^2 + 2\sigma_k^2 \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k) \\ = & \sum_{i=1}^m \left|\left| \sqrt{\tau_{ik}^{(\nu)}} \boldsymbol{I}_{n_i} \boldsymbol{y}_i - \sqrt{\tau_{ik}^{(\nu)}} \boldsymbol{I}_{n_i} \boldsymbol{X}_i \boldsymbol{\beta}_k - \sqrt{\tau_{ik}^{(\nu)}} \boldsymbol{I}_{n_i} \boldsymbol{Z}_i \boldsymbol{D}_k \boldsymbol{\Gamma}_k \boldsymbol{b}_{ki} \right|\right|^2 + 2\sigma_k^2 \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k) \\ = & \left|\left| \sqrt{\boldsymbol{\tau}_k^{(\nu)}} \boldsymbol{y} - \sqrt{\boldsymbol{\tau}_k^{(\nu)}} \boldsymbol{X} \boldsymbol{\beta}_k - \sqrt{\boldsymbol{\tau}_k^{(\nu)}} \boldsymbol{Z} \tilde{\boldsymbol{D}}_k \tilde{\boldsymbol{\Gamma}}_k \boldsymbol{b}_k \right|\right|^2 + 2\sigma_k^2 \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k), \end{aligned} \tag{4.9}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')'$, $\boldsymbol{X} = (\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m')'$. $\boldsymbol{Z}$ represents the block diagonal matrix of $\boldsymbol{Z}_i$, and $\tilde{\boldsymbol{D}}_k = \boldsymbol{I}_m \otimes \boldsymbol{D}_k$ and $\tilde{\boldsymbol{\Gamma}}_k = \boldsymbol{I}_m \otimes \boldsymbol{\Gamma}_k$, where $\otimes$ denotes the Kronecker product. $\sqrt{\boldsymbol{\tau}_k^{(\nu)}}$ represents the following diagonal matrix

$$\sqrt{\boldsymbol{\tau}_k^{(\nu)}} = \begin{bmatrix} \sqrt{\tau_{1k}^{(\nu)}} \boldsymbol{I}_{n_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\tau_{2k}^{(\nu)}} \boldsymbol{I}_{n_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\tau_{mk}^{(\nu)}} \boldsymbol{I}_{n_m} \end{bmatrix}.$$

If we define the transformed data in expression (4.9) as $\boldsymbol{y}_{(k)} = \sqrt{\boldsymbol{\tau}_k^{(\nu)}}\boldsymbol{y}$, $\boldsymbol{X}_{(k)} = \sqrt{\boldsymbol{\tau}_k^{(\nu)}}\boldsymbol{X}$, and $\boldsymbol{Z}_{(k)} = \sqrt{\boldsymbol{\tau}_k^{(\nu)}}\boldsymbol{Z}$, then it is as if we had a set of LME data $(\boldsymbol{y}_{(k)}, \boldsymbol{X}_{(k)}, \boldsymbol{Z}_{(k)})$. The subscript k emphasizes the fact that for each mixture component, we have a different set of transformed data due to the weights $\sqrt{\boldsymbol{\tau}_k^{(\nu)}}$. Equation (4.9) can thus be rewritten as

$$\left|\left|\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k - \boldsymbol{Z}_{(k)}\tilde{\boldsymbol{D}}_k\tilde{\boldsymbol{\Gamma}}_k\boldsymbol{b}_k\right|\right|^2 + 2\sigma_k^2 \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k), \tag{4.10}$$

of which we want to minimize the conditional expectation. Let ω index the inner EM iterations. Then, at iteration $\omega + 1$, our objective function is

$$Q(\boldsymbol{\phi}_k|\boldsymbol{\phi}_k^{(\omega)}) = \mathrm{E}_{\boldsymbol{b}_k|\boldsymbol{y},\boldsymbol{\phi}_k^{(\omega)}}\left\{\left|\left|\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k - \boldsymbol{Z}_{(k)}\tilde{\boldsymbol{D}}_k\tilde{\boldsymbol{\Gamma}}_k\boldsymbol{b}_k\right|\right|^2\right\} + 2\sigma_k^2 \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k). \tag{4.11}$$

Inner E-step:

In the E-step, we find the conditional distribution of $\boldsymbol{b}_k$ given $\boldsymbol{\phi}_k^{(\omega)}$ and $\boldsymbol{y}$. The joint density of $\boldsymbol{b}_{ki}$ and $\boldsymbol{y}_i$ for a single subject i was found in (4.6). Using the fact that both the observations and the random effects are independent between subjects, the joint density of $\boldsymbol{b}_k$ and $\boldsymbol{y}$ given current estimates $\boldsymbol{\phi}_k^{(\omega)}$ is

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{b}_k|\boldsymbol{\phi}_k^{(\omega)}) &= \prod_{i=1}^{m} \frac{1}{(2\pi\sigma_k^{2(\omega)})^{\frac{n_i+q}{2}}} \exp\left(-\frac{1}{2\sigma_k^{2(\omega)}}||\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k^{(\omega)} - \boldsymbol{Z}_i\boldsymbol{D}_k^{(\omega)}\boldsymbol{\Gamma}_k^{(\omega)}\boldsymbol{b}_{ki}||^2 - \frac{1}{2\sigma_k^{2(\omega)}}\boldsymbol{b}_{ki}'\boldsymbol{b}_{ki}\right) \\ &= \frac{1}{(2\pi\sigma_k^{2(\omega)})^{\frac{N+mq}{2}}} \exp\left(-\frac{1}{2\sigma_k^{2(\omega)}}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(\omega)} - \boldsymbol{Z}\tilde{\boldsymbol{D}}_k^{(\omega)}\tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\boldsymbol{b}_k||^2 - \frac{1}{2\sigma_k^{2(\omega)}}\boldsymbol{b}_k'\boldsymbol{b}_k\right). \end{aligned}$$

Therefore, the full conditional distribution of $\boldsymbol{b}_k$, given $\boldsymbol{\phi}_k^{(\omega)}$ and $\boldsymbol{y}$, is given by

$$\begin{aligned} f(\boldsymbol{b}_k|\boldsymbol{y}, \boldsymbol{\phi}_k^{(\omega)}) &\propto \exp\left\{-\frac{1}{2\sigma_k^{2(\omega)}}\Big(||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(\omega)} - \boldsymbol{Z}\tilde{\boldsymbol{D}}_k^{(\omega)}\tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\boldsymbol{b}_k||^2 + \boldsymbol{b}_k'\boldsymbol{b}_k\Big)\right\} \\ &= \exp\Bigg\{ -\frac{1}{2\sigma_k^{2(\omega)}}\Bigg[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(\omega)})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(\omega)}) - 2(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(\omega)})'\boldsymbol{Z}\tilde{\boldsymbol{D}}_k^{(\omega)}\tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\boldsymbol{b}_k \\ &\qquad + \boldsymbol{b}_k'\tilde{\boldsymbol{\Gamma}}_k'^{(\omega)}\tilde{\boldsymbol{D}}_k^{(\omega)}\boldsymbol{Z}'\boldsymbol{Z}\tilde{\boldsymbol{D}}_k^{(\omega)}\tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\boldsymbol{b}_k + \boldsymbol{b}_k'\boldsymbol{b}_k\Bigg]\Bigg\} \end{aligned}$$

$$\propto \exp\Bigg\{ -\frac{1}{2\sigma_k^{2(\omega)}} \Bigg[\boldsymbol{b}_k' \Big(\tilde{\boldsymbol{\Gamma}}_k'^{(\omega)} \tilde{\boldsymbol{D}}_k^{(\omega)} \boldsymbol{Z}'\boldsymbol{Z} \tilde{\boldsymbol{D}}_k^{(\omega)} \tilde{\boldsymbol{\Gamma}}_k^{(\omega)} + \boldsymbol{I}_{mq} \Big) \boldsymbol{b}_k \\ - 2\Big((\boldsymbol{Z}\tilde{\boldsymbol{D}}_k^{(\omega)} \tilde{\boldsymbol{\Gamma}}_k^{(\omega)})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(\omega)}) \Big)' \boldsymbol{b}_k \Bigg] \Bigg\}. \qquad (4.12)$$

After completing the square inside the exponential, (4.12) can be recognized as the kernel of a multivariate normal distribution. In the end, we find that $\boldsymbol{b}_k | \boldsymbol{y}, \boldsymbol{\phi}_k^{(\omega)} \sim \mathcal{N}(\hat{\boldsymbol{b}}_k^{(\omega)}, \boldsymbol{U}_k^{(\omega)})$, where the mean and variance are given by

$$\hat{\boldsymbol{b}}_k^{(\omega)} = (\tilde{\boldsymbol{\Gamma}}_k'^{(\omega)} \tilde{\boldsymbol{D}}_k^{(\omega)} \boldsymbol{Z}'\boldsymbol{Z} \tilde{\boldsymbol{D}}_k^{(\omega)} \tilde{\boldsymbol{\Gamma}}_k^{(\omega)} + \boldsymbol{I}_{mq})^{-1} (\boldsymbol{Z}\tilde{\boldsymbol{D}}_k^{(\omega)} \tilde{\boldsymbol{\Gamma}}_k^{(\omega)})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_k^{(\omega)}) \\ \text{and} \quad \boldsymbol{U}_k^{(\omega)} = \sigma_k^{2(\omega)} (\tilde{\boldsymbol{\Gamma}}_k'^{(\omega)} \tilde{\boldsymbol{D}}_k^{(\omega)} \boldsymbol{Z}'\boldsymbol{Z} \tilde{\boldsymbol{D}}_k^{(\omega)} \tilde{\boldsymbol{\Gamma}}_k^{(\omega)} + \boldsymbol{I}_{mq})^{-1}, \qquad (4.13)$$

respectively. This completes the inner E-step.

Inner M-step:

In the inner M-step, we minimize our objective function (4.11) with respect to the parameters $\boldsymbol{\phi}_k = (\boldsymbol{\beta}_k', \boldsymbol{d}_k', \boldsymbol{\gamma}_k', \sigma_k^2)'$. This optimization is done by iterating between $(\boldsymbol{\beta}_k', \boldsymbol{d}_k')'$, $\boldsymbol{\gamma}_k$, and σ_k^2 as follows.

First, we update $(\boldsymbol{\beta}_k', \boldsymbol{d}_k')'$. Bondell et al. (2010) showed that, omitting terms that do not involve $\boldsymbol{\phi}_k$, expression (4.10) can be rewritten as

$$\begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix}' \begin{bmatrix} \boldsymbol{X}_{(k)}'\boldsymbol{X}_{(k)} & \boldsymbol{X}_{(k)}'\boldsymbol{Z}_{(k)}\text{Diag}(\tilde{\boldsymbol{\Gamma}}_k\boldsymbol{b}_k)(\boldsymbol{1}_m \otimes \boldsymbol{I}_q) \\ (\boldsymbol{1}_m \otimes \boldsymbol{I}_q)'\text{Diag}(\tilde{\boldsymbol{\Gamma}}_k\boldsymbol{b}_k)\boldsymbol{Z}_{(k)}'\boldsymbol{X}_{(k)} & (\boldsymbol{1}_m \otimes \boldsymbol{I}_q)' \left(\boldsymbol{Z}_{(k)}'\boldsymbol{Z}_{(k)} \bullet \tilde{\boldsymbol{\Gamma}}_k \boldsymbol{b}_k \boldsymbol{b}_k' \tilde{\boldsymbol{\Gamma}}_k' \right) (\boldsymbol{1}_m \otimes \boldsymbol{I}_q) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix} \\ - 2\boldsymbol{y}_{(k)}' \begin{bmatrix} \boldsymbol{X}_{(k)} & \boldsymbol{Z}_{(k)}\text{Diag}(\tilde{\boldsymbol{\Gamma}}_k\boldsymbol{b}_k)(\boldsymbol{1}_m \otimes \boldsymbol{I}_q) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix} + 2\sigma_k^2 \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k), \qquad (4.14)$$

where $\bullet$ represents the Hadamard product, and $\otimes$ the Kronecker product. After taking the conditional expectation of (4.14), and fixing $\boldsymbol{\Gamma}_k$ and σ_k^2 at their current estimates, we obtain our penalized quadratic objective function for $(\boldsymbol{\beta}_k', \boldsymbol{d}_k')'$ as

$$
\begin{aligned}
&Q(\boldsymbol{\beta}_k, \boldsymbol{d}_k|\boldsymbol{\phi}_k^{(\omega)}) = \\
&\begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix}' \begin{bmatrix} \boldsymbol{X}_{(k)}'\boldsymbol{X}_{(k)} & \boldsymbol{X}_{(k)}'\boldsymbol{Z}_{(k)}\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\hat{\boldsymbol{b}}_k^{(\omega)})(\mathbf{1}_m \otimes \boldsymbol{I}_q) \\ (\mathbf{1}_m \otimes \boldsymbol{I}_q)'\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\hat{\boldsymbol{b}}_k^{(\omega)})\boldsymbol{Z}_{(k)}'\boldsymbol{X}_{(k)} & (\mathbf{1}_m \otimes \boldsymbol{I}_q)'\left(\boldsymbol{Z}_{(k)}'\boldsymbol{Z}_{(k)} \bullet \tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\hat{\boldsymbol{G}}_k^{(\omega)}\tilde{\boldsymbol{\Gamma}}_k^{(\omega)'}\right)(\mathbf{1}_m \otimes \boldsymbol{I}_q) \end{bmatrix} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} \\
&\qquad - 2\boldsymbol{y}_{(k)}' \begin{bmatrix}\boldsymbol{X}_{(k)} & \boldsymbol{Z}_{(k)}\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}_k^{(\omega)}\hat{\boldsymbol{b}}_k^{(\omega)})(\mathbf{1}_m \otimes \boldsymbol{I}_q)\end{bmatrix} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} + 2\sigma_k^{2(\omega)} \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k) \\
&= \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix}' \boldsymbol{M}^{(\omega)} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} - 2\boldsymbol{y}_{(k)}'\boldsymbol{L}^{(\omega)} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} + 2\sigma_k^{2(\omega)} \cdot \boldsymbol{p}_{mk}(\boldsymbol{\phi}_k), \qquad (4.15)
\end{aligned}
$$

where we simply replaced $\boldsymbol{b}_k$ and $\boldsymbol{b}_k\boldsymbol{b}_k'$ in equation (4.14) by their conditional expectations, $\hat{\boldsymbol{b}}_k^{(\omega)}$ and $\mathrm{E}\,(\boldsymbol{b}_k\boldsymbol{b}_k') = \hat{\boldsymbol{G}}_k^{(\omega)} = \boldsymbol{U}^{(\omega)} + \hat{\boldsymbol{b}}_k^{(\omega)}\hat{\boldsymbol{b}}_k^{(\omega)'}$, respectively.

In order to obtain closed form solution for (4.15), we adopt a local quadratic approximation proposed by Fan and Li (2001) and replace $p_{mk}(\theta)$ by

$$
p_{mk}(\theta_0) + \frac{p_m'(|\theta_0|)}{2|\theta_0|}(\theta^2 - \theta_0^2)
$$

in a neighborhood of θ_0. Equation (4.15) can thus be locally approximated (except for a constant term) by

$$
\begin{aligned}
Q(\boldsymbol{\beta}_k, \boldsymbol{d}_k|\boldsymbol{\phi}_k^{(\omega)}) &\approx \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix}' \boldsymbol{M}^{(\omega)} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} - 2\boldsymbol{y}_{(k)}'\boldsymbol{L}^{(\omega)} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} + \frac{2\sigma_k^{2(\omega)} \cdot \pi_k^{(\nu+1)} \cdot m}{2} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix}' \boldsymbol{\Sigma}_m(\boldsymbol{\phi}_k^{(\omega)}) \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} \\
&= \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix}' \left(\boldsymbol{M}^{(\omega)} + \sigma_k^{2(\omega)} \cdot \pi_k^{(\nu+1)} \cdot m \cdot \boldsymbol{\Sigma}_m(\boldsymbol{\phi}_k^{(\omega)})\right) \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix} - 2\boldsymbol{y}_{(k)}'\boldsymbol{L}^{(\omega)} \begin{bmatrix}\boldsymbol{\beta}_k \\ \boldsymbol{d}_k\end{bmatrix}, \qquad (4.16)
\end{aligned}
$$

where $\boldsymbol{\Sigma}_m(\boldsymbol{\phi}_k^{(\omega)}) = \mathrm{diag}\left\{\frac{p_m'(|\beta_{k1}^{(\omega)}|)}{|\beta_{k1}^{(\omega)}|}, \ldots, \frac{p_m'(|\beta_{kp}^{(\omega)}|)}{|\beta_{kp}^{(\omega)}|}, \frac{p_m'(|d_{k1}^{(\omega)}|)}{|d_{k1}^{(\omega)}|}, \ldots, \frac{p_m'(|d_{kq}^{(\omega)}|)}{|d_{kq}^{(\omega)}|}\right\}$. Since d_{kj}'s are nonnegative for all $j = 1, \ldots, q$ and $k = 1, \ldots, K$, the updated $(\boldsymbol{\beta}_k', \boldsymbol{d}_k')$ is given by the solution to the following constrained quadratic minimization

$$\begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix}^{(\omega+1)} = \arg\max \begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix}' \left(\boldsymbol{M}^{(\omega)} + \sigma_k^{2(\omega)} \cdot \pi_k^{(\nu+1)} \cdot m \cdot \boldsymbol{\Sigma}_m(\boldsymbol{\phi}_k^{(\omega)})\right) \begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix} - 2\boldsymbol{y}'_{(k)}\boldsymbol{L}^{(\omega)} \begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix},$$
$$\text{subject to } d_{kj} \geq 0 \text{ for } j = 1, \ldots, q. \tag{4.17}$$

Quadratic programming techniques could be used to solve (4.17). Instead, in our simulations, we updated $(\boldsymbol{\beta}'_k, \boldsymbol{d}'_k)'$ with the solution to the derivative function of (4.16), namely

$$\begin{bmatrix} \boldsymbol{\beta}_k \\ \boldsymbol{d}_k \end{bmatrix}^{(\omega+1)} = \left(\boldsymbol{M}^{(\omega)} + \sigma_k^{2(\omega)} \cdot \pi_k^{(\nu+1)} \cdot m \cdot \boldsymbol{\Sigma}_m(\boldsymbol{\phi}_k^{(\omega)})\right)^{-1} \cdot (\boldsymbol{L}^{(\omega)\prime}\boldsymbol{y}_{(k)}), \tag{4.18}$$

and set to zero any negative estimate of d_{kj}, since the latter are the square roots of the diagonal elements of the Cholesky decomposition of the random-effect variance matrix. Such techniques of projection onto the space of nonnegative definite matrices are common in estimating the variance components in LME models (Demidenko, 2004). This worked well in our simulations, as none of the truly nonzero d_{kj}'s was ever estimated to be negative. Also, when any truly zero d_{kj} had a negative estimate, it was extremely close to zero.

If d_{kl} is 0, then the γ's in the l-th row of $\boldsymbol{\Gamma}_k$ are automatically set to 0. This is necessary to ensure the identifiability of the γ's. To update $\boldsymbol{\gamma}_k$, we first rewrite the squared vector norm $\left|\left|\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k - \boldsymbol{Z}_{(k)}\tilde{\boldsymbol{D}}_k\tilde{\boldsymbol{\Gamma}}_k\boldsymbol{b}_k\right|\right|^2$ in a quadratic form for $\boldsymbol{\gamma}_k$ and then compute its conditional expectation. Bondell et al. (2010) showed that the objective function for $\boldsymbol{\gamma}_k$ is given by

$$Q(\boldsymbol{\gamma}_k|\boldsymbol{\phi}_k^{(\omega)}) = \boldsymbol{\gamma}'_k\boldsymbol{P}^{(\omega)}\boldsymbol{\gamma} - 2\left\{\left(\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k^{(\omega)}\right)'\boldsymbol{R}^{(\omega)} - \boldsymbol{T}^{(\omega)\prime}\right\}\boldsymbol{\gamma}_k, \tag{4.19}$$

for some matrices $\boldsymbol{P}^{(\omega)}$, $\boldsymbol{T}^{(\omega)}$ and $\boldsymbol{R}^{(\omega)}$. The minimizer of (4.19) is given by

$$\boldsymbol{\gamma}_k^{(\omega+1)} = \left(\boldsymbol{P}^{(\omega)}\right)^{-}\left\{\boldsymbol{R}^{(\omega)\prime}\left(\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k^{(\omega)}\right) - \boldsymbol{T}^{(\omega)\prime}\right\}, \tag{4.20}$$

where $\left(\boldsymbol{P}^{(\omega)}\right)^{-}$ denotes the Moore-Penrose generalized inverse of $\boldsymbol{P}^{(\omega)}$.

We update σ_k^2 by directly maximizing equation (4.5) with respect to σ_k^2. By solving its first derivative function, we find that the maximizer is given by

$$\sigma_k^2 = \frac{\sum_{i=1}^m \tau_{ik}^{(\nu)} \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k\right)' \left(\boldsymbol{Z}_i\boldsymbol{D}_k\boldsymbol{\Gamma}_k\boldsymbol{\Gamma}_k'\boldsymbol{D}_k\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i}\right)^{-1} \left(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_k\right)}{\sum_{i=1}^m \tau_{ik}^{(\nu)} \cdot n_i}.$$

If we factor the weights $\tau_{ik}^{(\nu)}$ into the residuals, then we can rewrite the above equation in terms of the transformed data $(\boldsymbol{y}_{(k)}, \boldsymbol{X}_{(k)}, \boldsymbol{Z}_{(k)})$ as

$$\sigma_k^2 = \frac{\left(\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k\right)' \left(\boldsymbol{Z}_{(k)}\tilde{\boldsymbol{D}}_k\tilde{\boldsymbol{\Gamma}}_k\tilde{\boldsymbol{\Gamma}}_k'\tilde{\boldsymbol{D}}_k\boldsymbol{Z}_{(k)}' + \boldsymbol{I}_N\right)^{-1} \left(\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k\right)}{\sum_{i=1}^m \tau_{ik}^{(\nu)} \cdot n_i}. \tag{4.21}$$

Observe that the numerator is a weighted sum of squared residuals, and the denominator is the estimated virtual sample size of the k-th component.

We assume that the current estimates $(\boldsymbol{\beta}_k^{(\omega+1)'}, \boldsymbol{d}_k^{(\omega+1)'}, \boldsymbol{\gamma}_k^{(\omega+1)'})'$ are the true values for these parameters, so we replace $(\boldsymbol{\beta}_k', \boldsymbol{d}_k', \boldsymbol{\gamma}_k')'$ in (4.21) by their current estimates to update σ_k^2, i.e.

$$\sigma_k^{2(\omega+1)} = \frac{\left(\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k^{(\omega+1)}\right)' \left(\boldsymbol{Z}_{(k)}\tilde{\boldsymbol{D}}_k^{(\omega+1)}\tilde{\boldsymbol{\Gamma}}_k^{(\omega+1)}\tilde{\boldsymbol{\Gamma}}_k^{(\omega+1)'}\tilde{\boldsymbol{D}}_k^{(\omega+1)}\boldsymbol{Z}_{(k)}' + \boldsymbol{I}_N\right)^{-1} \left(\boldsymbol{y}_{(k)} - \boldsymbol{X}_{(k)}\boldsymbol{\beta}_k^{(\omega+1)}\right)}{\sum_{i=1}^m \tau_{ik}^{(\nu)} \cdot n_i}.$$

Once the parameters $(\boldsymbol{\beta}_k', \boldsymbol{d}_k', \boldsymbol{\gamma}_k', \sigma_k^2)'$ have converged in the inner EM algorithm, say to the values $(\boldsymbol{\beta}_k^{(\nu+1)'}, \boldsymbol{d}_k^{(\nu+1)'}, \boldsymbol{\gamma}_k^{(\nu+1)'}, \sigma_k^{2(\nu+1)})'$, the maximization of equation (4.5) is complete. We can then update $\boldsymbol{\phi}_k$ in the outer EM algorithm by

$$\boldsymbol{\phi}_k^{(\nu+1)} = \begin{bmatrix} \boldsymbol{\beta}_k^{(\nu+1)} \\ \boldsymbol{d}_k^{(\nu+1)} \\ \boldsymbol{\gamma}_k^{(\nu+1)} \\ \sigma_k^{2(\nu+1)} \end{bmatrix}.$$

We iterate between the outer E-step and M-step until all parameters in $\boldsymbol{\Phi}$ have reached convergence.

4.1 Choice of tuning parameters

When using the LASSO, Adaptive LASSO or SCAD penalties, the penalty function for each mixture component k would involve a tuning parameter λ_{mk}. The nested EM algorithm described above applies to a fixed tuning parameter vector $\boldsymbol{\lambda}_m = (\lambda_{m1}, \ldots, \lambda_{mK})$. In practice, we need to choose each λ_{mk} from a grid of candidate values.

For an LME model, Bondell et al. (2010) proposed to use the following BIC-type criterion,

$$BIC(\lambda_m) = -2l(\hat{\boldsymbol{\phi}}) + \log(N) \times df_{\lambda_m}, \tag{4.22}$$

where $\hat{\boldsymbol{\phi}}$ is the vector of maximum penalized likelihood estimates for all LME parameters, l denotes the log-likelihood, N is the total sample size, and df_{λ_m} is the number of non-zero coefficients in $\hat{\boldsymbol{\phi}}$. We make suitable changes to this BIC criterion to reflect our mixture structure.

To do this, we first find the ML estimates, $\hat{\boldsymbol{\Phi}}^{(ml)}$, for the FMLME model (this is done using the same nested EM algorithm described before, but without the penalty). We then calculate the posterior probability of observation $\boldsymbol{y}_i$ belonging to the j-th component, namely

$$\tau_{ij}^{(ml)} = \frac{\hat{\pi}_j^{(ml)} f_j(\boldsymbol{y}_i; \hat{\boldsymbol{\phi}}_j^{(ml)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(ml)} f_k(\boldsymbol{y}_i; \hat{\boldsymbol{\phi}}_k^{(ml)})}, \tag{4.23}$$

for each subject i and each component j.

Suppose we are choosing the tuning parameter for component k. We define the following weight matrix,

$$\sqrt{\boldsymbol{\tau}_k^{(ml)}} = \begin{bmatrix} \sqrt{\tau_{1k}^{(ml)}}\boldsymbol{I}_{n_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\tau_{2k}^{(ml)}}\boldsymbol{I}_{n_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\tau_{mk}^{(ml)}}\boldsymbol{I}_{n_m} \end{bmatrix}.$$

Analogously to the previous section, we define $\boldsymbol{y}_{(k)}^{(ml)} = \sqrt{\boldsymbol{\tau}_k^{(ml)}}\boldsymbol{y}$, $\boldsymbol{X}_{(k)}^{(ml)} = \sqrt{\boldsymbol{\tau}_k^{(ml)}}\boldsymbol{X}$, and $\boldsymbol{Z}_{(k)}^{(ml)} = \sqrt{\boldsymbol{\tau}_k^{(ml)}}\boldsymbol{Z}$. We can again consider $(\boldsymbol{y}_{(k)}^{(ml)}, \boldsymbol{X}_{(k)}^{(ml)}, \boldsymbol{Z}_{(k)}^{(ml)})$ as a set of LME data, and use the methodology of Bondell et al. (2010) to find the maximum penalized likelihood estimate, $\hat{\boldsymbol{\phi}}_k^{(mpl)}$, of $\boldsymbol{\phi}_k$.

Let $N_k = \sum_{i=1}^m \tau_{ik}^{(ml)} \cdot n_i$ be the estimated virtual sample size of the k-th component, with $k = (1, \ldots, K)$. Similar to the BIC-2 of Steele and Raftery (2010), our component-wise BIC criterion for the FMLME model is then defined as

$$BIC_k(\lambda_{mk}) = -2l_k(\hat{\boldsymbol{\phi}}_k^{(mpl)}) + \log(N_k) \times df_{\lambda_{mk}}, \qquad (k = 1, \ldots, K), \tag{4.24}$$

where $l_k(\hat{\boldsymbol{\phi}}_k^{(mpl)}) = \sum_{i=1}^m \tau_{ik}^{(ml)} \log f_k(y_i; \hat{\boldsymbol{\phi}}_k^{(mpl)})$ is the weighted log-likelihood of the k-th component. This definition is in line with (4.22) and the modifications take into account the mixture structure. The tuning parameters λ_{mk} are chosen one at a time by minimizing $BIC_k(\lambda_{mk})$.

This concludes our numerical algorithm for fixed and random effects selection and estimation. In the next chapter, we examine the asymptotic properties of our penalized estimators.

CHAPTER 5
Asymptotic properties

Suppose that the data $\{(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) : i = 1, \ldots, m\}$ are a random sample from the FMLME model (3.1). Write $\boldsymbol{\Phi} = (\phi_1, \phi_2, \ldots, \phi_T)$ so that T is the total number of parameters in the model. In order to study the asymptotic properties of the proposed methodology, several regularity conditions have to be imposed on the joint distribution of $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$. Let $f(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\Phi})$ be the joint density function of $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $\boldsymbol{\Omega}$ be an open parameter space. We assume the following regularity conditions are satisfied.

Regularity Conditions:

A_1 $f(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\Phi})$ is identifiable in $\boldsymbol{\Phi}$ up to a permutation of the components of the mixture.

A_2 For each $\boldsymbol{\Phi}_0 \in \boldsymbol{\Omega}$, there exist $M_{1i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, $M_{2i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ and $M_{3i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ (possibly depending on $\boldsymbol{\Phi}_0$) such that for $\boldsymbol{\Phi}$ in a neighborhood of $\boldsymbol{\Phi}_0$,

$$\left|\frac{\partial \log f(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\Phi})}{\partial \phi_j}\right| < M_{1i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i),$$
$$\left|\frac{\partial^2 \log f(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\Phi})}{\partial \phi_j \partial \phi_l}\right| < M_{2i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i),$$
$$\left|\frac{\partial^3 \log f(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\Phi})}{\partial \phi_j \partial \phi_l \partial \phi_n}\right| < M_{3i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i),$$

such that $\mathrm{E}_{\boldsymbol{\Phi}_0}(M_{1i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)) < \infty$, $\mathrm{E}_{\boldsymbol{\Phi}_0}(M_{2i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)) < \infty$ and $\mathrm{E}_{\boldsymbol{\Phi}_0}(M_{3i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i))$ $< \infty$.

A_3 The Fisher information matrix $I(\boldsymbol{\Phi})$ is finite and positive definite for all $\boldsymbol{\Phi} \in \boldsymbol{\Omega}$.

Note that condition A_2 is on the joint density function of $(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ only, which does not involve a penalty function, therefore the previously discussed non-differentiability of LASSO and SCAD functions does not affect the feasibility of this assumption.

Also note that even though the density of FMLME model (3.1) is a finite mixture of multivariate normal distributions, in this chapter, the asymptotic properties are determined in greater generality, i.e. the component-wise densities are not necessarily multivariate normals, but the joint density $f(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i; \boldsymbol{\Phi})$ satisfies the conditions A_1–A_3.

Decompose the parameter $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1', \boldsymbol{\Phi}_2')'$ such that $\boldsymbol{\Phi}_2$ contains all zero effects from all the mixture components, and split the vector of true parameter values accordingly as $\boldsymbol{\Phi}_0 = (\boldsymbol{\Phi}_{10}', \boldsymbol{\Phi}_{20}')'$. Denote the elements of $\boldsymbol{\Phi}_{10}$ with a superscript such as β_{kj}^{10} and d_{kj}^{10}. Our asymptotic results involve the following quantities:

$$\begin{aligned} a_m &= \max_{k,j} \left\{ \sqrt{m} \left| p_{mk}(\beta_{kj}^{10}) \right|, \sqrt{m} \left| p_{mk}(d_{kj}^{10}) \right| \right\}, \\ b_m &= \max_{k,j} \left\{ \sqrt{m} \left| p_{mk}'(\beta_{kj}^{10}) \right|, \sqrt{m} \left| p_{mk}'(d_{kj}^{10}) \right| \right\}, \\ \text{and} \quad c_m &= \max_{k,j} \left\{ \sqrt{m} \left| p_{mk}''(\beta_{kj}^{10}) \right|, \sqrt{m} \left| p_{mk}''(d_{kj}^{10}) \right| \right\}. \end{aligned}$$

Assume that the penalty functions p_{mk} satisfy the following conditions:

P_0 For all m and k, $p_{mk}(0) = 0$ and $p_{mk}(.)$ is symmetric and non-negative. In addition, it is non-decreasing and twice differentiable on $(0, \infty)$ with at most a few exceptions.

P_1 As $m \to \infty$, $a_m = o(1 + b_m)$, $c_m = o(\sqrt{m})$.

P_2 For $N_n = \{\theta : 0 < \theta \leq m^{-1/2} \log m\}$, $\lim_{m \to \infty} \inf_{\theta \in N_n} \sqrt{m} p_{mk}'(\theta) = \infty$.

Our penalized likelihood estimator possesses the following asymptotic properties.

Theorem 1. *Let* $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1', 0')'$, *and the observations follow the FMLME model* (3.1) *satisfying regularity conditions* $A_1 - A_3$, *and assume the penalty function* p_{mk} *satisfies* P_0 *and* P_1. *Then there exists a local maximizer* $\hat{\boldsymbol{\Phi}}_m$ *of the penalized log-likelihood function* (3.3) *such that* $||\hat{\boldsymbol{\Phi}}_m - \boldsymbol{\Phi}_0|| = O_p\left\{m^{-1/2}(1 + b_m)\right\}$.

Proof. Let $r_m = m^{-1/2}(1 + b_m)$. It suffices to show that for any small enough $\varepsilon > 0$, there exists a constant M_ε such that for sufficiently large m,

$$\Pr\left\{ \sup_{||\boldsymbol{u}||=M_\varepsilon} pl_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) < pl_m(\boldsymbol{\Phi}_0) \right\} \geq 1 - \varepsilon.$$

So with large probability, there exists a local maximum in $\{\boldsymbol{\Phi}_0 + r_m\boldsymbol{u} : ||\boldsymbol{u}|| \leq M_\varepsilon\}$. This local maximizer $\hat{\boldsymbol{\Phi}}_m$ satisfies $||\hat{\boldsymbol{\Phi}}_m - \boldsymbol{\Phi}_0|| = O_p\left\{m^{-1/2}(1 + b_m)\right\}$.

Let

$$\begin{aligned}\Delta_m(\boldsymbol{u}) &= pl_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - pl_m(\boldsymbol{\Phi}_0)\\ &= \{l_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - l_m(\boldsymbol{\Phi}_0)\} - \{\boldsymbol{p}_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - \boldsymbol{p}_m(\boldsymbol{\Phi}_0)\}.\end{aligned}$$

By assumption P_0, $\boldsymbol{p}_{mk}(0) = 0$, therefore $\boldsymbol{p}_m(\boldsymbol{\Phi}_0) = \boldsymbol{p}_m(\boldsymbol{\Phi}_{10})$. Given that $\boldsymbol{p}_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u})$ is a sum of positive terms, removing terms corresponding to zero components makes it smaller, hence

$$\begin{aligned}\Delta_m(\boldsymbol{u}) &\leq \{l_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - l_m(\boldsymbol{\Phi}_0)\} - \{\boldsymbol{p}_m(\boldsymbol{\Phi}_{10} + r_m\boldsymbol{u}_I) - \boldsymbol{p}_m(\boldsymbol{\Phi}_{10})\}\\ &\leq \{l_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - l_m(\boldsymbol{\Phi}_0)\} + |\boldsymbol{p}_m(\boldsymbol{\Phi}_{10} + r_m\boldsymbol{u}_I) - \boldsymbol{p}_m(\boldsymbol{\Phi}_{10})|, \qquad (5.1)\end{aligned}$$

where $\boldsymbol{u}_I$ is the sub-vector of $\boldsymbol{u}$ that corresponds to the non-zero effects. By Taylor's expansion,

$$\begin{aligned}l_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - l_m(\boldsymbol{\Phi}_0) &= r_m l_m'(\boldsymbol{\Phi}_0)^T\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T(l_m''(\boldsymbol{\Phi}_0))\boldsymbol{u}r_m^2\\ &= \frac{(1 + b_m)}{\sqrt{m}} l_m'(\boldsymbol{\Phi}_0)^T\boldsymbol{u} + \frac{(1 + b_m)^2}{2m}\boldsymbol{u}^T(l_m''(\boldsymbol{\Phi}_0))\boldsymbol{u}, \qquad (5.2)\end{aligned}$$

where we omitted the remainder term, since it vanishes as $m \to \infty$ by regularity condition A_2. For the hessian matrix $l_m''(\boldsymbol{\Phi}_0)$, we have

$$\frac{1}{m} l_m''(\boldsymbol{\Phi}_0) \xrightarrow{p} -I(\boldsymbol{\Phi}_0).$$

Therefore,

$$\begin{aligned} l_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - l_m(\boldsymbol{\Phi}_0) &= \frac{(1+b_m)}{\sqrt{m}} l_m'(\boldsymbol{\Phi}_0)^T\boldsymbol{u} - \frac{(1+b_m)^2}{2}\boldsymbol{u}^T I(\boldsymbol{\Phi}_0)\boldsymbol{u}\{1+o_p(1)\} \\ &= (1+b_m)O_p(1)\,||\boldsymbol{u}|| - \frac{(1+b_m)^2}{2}\boldsymbol{u}^T I(\boldsymbol{\Phi}_0)\boldsymbol{u}\{1+o_p(1)\}, \qquad (5.3) \end{aligned}$$

since $\frac{1}{\sqrt{m}} l_m'(\boldsymbol{\Phi}_0) = O_p(1)$ by regularity conditions. Also, by regularity condition A_3, $I(\boldsymbol{\Phi}_0)$ is positive definite, therefore its smallest eigenvalue $\eta_{\min}$ is positive. Furthermore, we have that $\boldsymbol{u}^T I(\boldsymbol{\Phi}_0)\boldsymbol{u} \geq \eta_{\min}\,||\boldsymbol{u}||^2$. Applying this result to (5.3), we have

$$l_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - l_m(\boldsymbol{\Phi}_0) \leq (1+b_m)O_p(1)\,||\boldsymbol{u}|| - \frac{(1+b_m)^2}{2}\eta_{\min}\,||\boldsymbol{u}||^2\{1+o_p(1)\}. \qquad (5.4)$$

On the other hand, by Taylor's expansion and the triangular inequality,

$$\begin{aligned} &|\boldsymbol{p}_m(\boldsymbol{\Phi}_{10} + r_m\boldsymbol{u}_I) - \boldsymbol{p}_m(\boldsymbol{\Phi}_{10})| \\ =&\boldsymbol{p}_m'(\boldsymbol{\Phi}_{10})^T r_m\boldsymbol{u}_I + \frac{r_m^2}{2}\boldsymbol{u}_I^T\boldsymbol{p}_m''(\boldsymbol{\Phi}_{10})\boldsymbol{u}_I\{1+o(1)\} \\ \leq& r_m\left|\boldsymbol{p}_m'(\boldsymbol{\Phi}_{10})^T\boldsymbol{u}_I\right| + \frac{r_m^2}{2}\left|\boldsymbol{u}_I^T\boldsymbol{p}_m''(\boldsymbol{\Phi}_{10})\boldsymbol{u}_I\right|\{1+o(1)\} \\ \leq& r_m\left|\left|\boldsymbol{p}_m'(\boldsymbol{\Phi}_{10})^T\right|\right|\cdot||\boldsymbol{u}_I|| + \frac{r_m^2}{2}||\mathrm{diag}(\boldsymbol{p}_m''(\boldsymbol{\Phi}_{10}))||\cdot||\boldsymbol{u}_I||^2\{1+o(1)\}. \qquad (5.5) \end{aligned}$$

Let t_k be the total number of true non-zero fixed and random effects in the k-th component, and let $t = \max\{t_k, k = 1, \ldots, K\}$. Let $\boldsymbol{\beta}^{10}$ and $\boldsymbol{d}^{10}$ denote respectively the vectors of β_{kj}^{10}'s and d_{kj}^{10}'s from all the components. We notice that for the first term of (5.5),

$$||\boldsymbol{p}_m'(\boldsymbol{\Phi}_{10})|| = ||\boldsymbol{p}_m'(\pi_1, \ldots, \pi_K)|| + \left|\left|\boldsymbol{p}_m'(\boldsymbol{\beta}^{10}, \boldsymbol{d}^{10})\right|\right|. \qquad (5.6)$$

Recall that $\boldsymbol{p}_m(\boldsymbol{\Phi}) = \sum_{k=1}^K \pi_k \cdot m\left(\sum_{j=1}^p p_{mk}(\beta_{kj}) + \sum_{j=1}^q p_{mk}(d_{kj})\right)$, therefore

$$\boldsymbol{p}_m'(\pi_1, \ldots, \pi_K) = \begin{bmatrix} m\left(\sum_{j=1}^p p_{mk}(\beta_{1j}^{10}) + \sum_{j=1}^q p_{mk}(d_{1j}^{10})\right) \\ \vdots \\ m\left(\sum_{j=1}^p p_{mk}(\beta_{Kj}^{10}) + \sum_{j=1}^q p_{mk}(d_{Kj}^{10})\right) \end{bmatrix}.$$

Hence,

$$||\boldsymbol{p}'_m(\pi_1,\ldots,\pi_K)|| = m\cdot\sqrt{\sum_{k=1}^{K}\left[\sum_{j=1}^{p}p_{mk}(\beta_{kj}^{10})+\sum_{j=1}^{q}p_{mk}(d_{kj}^{10})\right]^2} \le m\sqrt{\sum_{k=1}^{K}\left[t_k\cdot\frac{a_m}{\sqrt{m}}\right]^2}$$
$$= a_m\sqrt{m}\sqrt{\sum_{k=1}^{K}t_k^2} \;\le\; a_m\sqrt{m}\sqrt{\sum_{k=1}^{K}t^2} \;=\; a_m\sqrt{m}\sqrt{K}t.$$

Furthermore,

$$\begin{aligned}
\big|\big|\boldsymbol{p}'_m(\boldsymbol{\beta}^{10},\boldsymbol{d}^{10})\big|\big| &= \big|\big|\nabla\boldsymbol{p}_m(\beta_{11}^{10},\ldots,\beta_{1p}^{10},\ldots,\beta_{K1}^{10},\ldots,\beta_{Kp}^{10},d_{11}^{10},\ldots,d_{1q}^{10},\ldots,d_{K1}^{10},\ldots,d_{Kq}^{10})\big|\big| \\
&= m||\pi_1p'_{m1}(\beta_{11}^{10}),\ldots,\pi_1p'_{m1}(\beta_{1p}^{10}),\ldots,\pi_Kp'_{mK}(\beta_{K1}^{10}),\ldots,\pi_Kp'_{mK}(\beta_{Kp}^{10}), \\
&\qquad \pi_1p'_{m1}(d_{11}^{10}),\ldots,\pi_1p'_{m1}(d_{1q}^{10}),\ldots,\pi_Kp'_{mK}(d_{K1}^{10}),\ldots,\pi_Kp'_{mK}(d_{Kq}^{10})|| \\
&\le m||p'_{m1}(\beta_{11}^{10}),\ldots,p'_{m1}(\beta_{1p}^{10}),\ldots,p'_{mK}(\beta_{K1}^{10}),\ldots,p'_{mK}(\beta_{Kp}^{10}), \\
&\qquad p'_{m1}(d_{11}^{10}),\ldots,p'_{m1}(d_{1q}^{10}),\ldots,p'_{mK}(d_{K1}^{10}),\ldots,p'_{mK}(d_{Kq}^{10})|| \\
&= m\sqrt{\sum_{k=1}^{K}\sum_{j=1}^{p}p'_{mk}(\beta_{kj}^{10})^2+\sum_{k=1}^{K}\sum_{j=1}^{q}p'_{mk}(d_{kj}^{10})^2} \\
&= m\sqrt{\sum_{k=1}^{K}\left(\sum_{j=1}^{p}p'_{mk}(\beta_{kj}^{10})^2+\sum_{j=1}^{q}p'_{mk}(d_{kj}^{10})^2\right)} \\
&\le m\sqrt{\sum_{k=1}^{K}t_k\cdot\left(\frac{b_m}{\sqrt{m}}\right)^2} = \sqrt{m}b_m\sqrt{\sum_{k=1}^{K}t_k} \\
&\le b_m\sqrt{m}\sqrt{K\cdot t}.
\end{aligned}$$

Therefore,

$$||\boldsymbol{p}'_m(\boldsymbol{\Phi}_0)|| \le a_m\sqrt{m}\sqrt{K}t + b_m\sqrt{m}\sqrt{K\cdot t}.$$

Using (5.5), this leads to

$$\begin{aligned}
&|\boldsymbol{p}_m(\boldsymbol{\Phi}_{10} + r_m\boldsymbol{u}_I) - \boldsymbol{p}_m(\boldsymbol{\Phi}_{10})| \\
&\leq r_m \left(a_m\sqrt{m}\sqrt{K}t + b_m\sqrt{m}\sqrt{K\cdot t}\right) ||\boldsymbol{u}|| + \frac{r_m^2}{2} \left|\left|\mathrm{diag}(\boldsymbol{p}_m''(\boldsymbol{\Phi}_{10}))\right|\right| \cdot ||\boldsymbol{u}_I||^2 \{1 + o(1)\} \\
&= \frac{1+b_m}{\sqrt{m}} \left(a_m\sqrt{m}\sqrt{K}t + b_m\sqrt{m}\sqrt{K\cdot t}\right) ||\boldsymbol{u}|| + \frac{1}{2}\frac{(1+b_m)^2}{m} \left|\left|\mathrm{diag}(\boldsymbol{p}_m''(\boldsymbol{\Phi}_{10}))\right|\right| \cdot ||\boldsymbol{u}_I||^2 \{1 + o(1)\} \\
&= a_m(1+b_m)\sqrt{K}t\, ||\boldsymbol{u}|| + b_m(1+b_m)\sqrt{K\cdot t}\, ||\boldsymbol{u}|| + \frac{1}{2}\frac{(1+b_m)^2}{m} \left|\left|\mathrm{diag}(\boldsymbol{p}_m''(\boldsymbol{\Phi}_{10}))\right|\right| \cdot ||\boldsymbol{u}_I||^2 \{1 + o(1)\}.
\end{aligned} \tag{5.7}$$

Furthermore,

$$\begin{aligned}
||\mathrm{diag}(\boldsymbol{p}_m''(\boldsymbol{\Phi}_{10}))|| &= m\sqrt{\sum_{k=1}^{K}\sum_{j=1}^{p} p_{mk}''(\beta_{kj}^0)^2\pi_k^2 + \sum_{k=1}^{K}\sum_{j=1}^{q} p_{mk}''(d_{kj}^0)^2\pi_k^2} \\
&\leq m\sqrt{\sum_{k=1}^{K}\left(\sum_{j=1}^{p} p_{mk}'(\beta_{kj}^{10})^2 + \sum_{j=1}^{q} p_{mk}'(d_{kj}^{10})^2\right)} \\
&\leq m\sqrt{\sum_{k=1}^{K} t_k \cdot \left(\frac{c_m}{\sqrt{m}}\right)^2} = \sqrt{m}c_m\sqrt{\sum_{k=1}^{K} t_k} \\
&\leq c_m\sqrt{m}\sqrt{K\cdot t}.
\end{aligned}$$

Combining this result with (5.7) gives

$$\begin{aligned}
|\boldsymbol{p}_m(\boldsymbol{\Phi}_{10} + r_m\boldsymbol{u}_I) - \boldsymbol{p}_m(\boldsymbol{\Phi}_{10})| \leq a_m(1+b_m)\sqrt{K}t\, ||\boldsymbol{u}|| &+ b_m(1+b_m)\sqrt{K\cdot t}\, ||\boldsymbol{u}|| \\
&+ \frac{1}{2}\frac{(1+b_m)^2}{\sqrt{m}} c_m\sqrt{K\cdot t}\, ||\boldsymbol{u}||^2 (1 + o(1)).
\end{aligned} \tag{5.8}$$

Combining (5.1), (5.4) and (5.8), we have

$$\begin{aligned}
\Delta_m(\boldsymbol{u}) \leq (1+b_m)O_p(1)\, ||\boldsymbol{u}|| - \frac{(1+b_m)^2}{2}\eta_{\min}\, ||\boldsymbol{u}||^2 \{1 + o_p(1)\} + a_m(1+b_m)\sqrt{K}t\, ||\boldsymbol{u}|| \\
+ b_m(1+b_m)\sqrt{K\cdot t}\, ||\boldsymbol{u}|| + \frac{1}{2}\frac{(1+b_m)^2}{\sqrt{m}} c_m\sqrt{K\cdot t}\, ||\boldsymbol{u}||^2 (1 + o(1)).
\end{aligned}$$

Dividing both sides of the above inequality by $(1+b_m)^2$, we have

$$\begin{aligned}\frac{\Delta_m(\boldsymbol{u})}{(1+b_m)^2} \leq{}& \frac{1}{1+b_m} O_p(1)\,||\boldsymbol{u}|| - \frac{1}{2}\eta_{\min}\,||\boldsymbol{u}||^2\,\{1+o_p(1)\} \\ &+ \frac{a_m}{1+b_m}\sqrt{K}t\,||\boldsymbol{u}|| + \frac{b_m}{1+b_m}\sqrt{K\cdot t}\,||\boldsymbol{u}|| + \frac{1}{2}\frac{c_m}{\sqrt{m}}\sqrt{K\cdot t}\,||\boldsymbol{u}||^2\,(1+o(1)).\end{aligned}$$

By condition P_1, $a_m = o(1+b_m)$ and $c_m = o(\sqrt{m})$. Applying these conditions to the second row of the above inequality, we have

$$\frac{\Delta_m(\boldsymbol{u})}{(1+b_m)^2} \leq \frac{1}{1+b_m} O_p(1)\,||\boldsymbol{u}|| - \frac{1}{2}\eta_{\min}\,||\boldsymbol{u}||^2\,\{1+o_p(1)\} + \frac{b_m}{1+b_m}\sqrt{K\cdot t}\,||\boldsymbol{u}|| + o(1). \quad (5.9)$$

Using (5.9), we have

$$\begin{aligned}&\Pr\left\{\sup_{||\boldsymbol{u}||=M_\varepsilon} \Delta_m(\boldsymbol{u}) < 0\right\} = \Pr\left\{\sup_{||\boldsymbol{u}||=M_\varepsilon} \frac{\Delta_m(\boldsymbol{u})}{(1+b_m)^2} < 0\right\} \\ \geq{}& \Pr\left\{\sup_{||\boldsymbol{u}||=M_\varepsilon}\left[\frac{1}{1+b_m}O_p(1)\,||\boldsymbol{u}|| - \frac{1}{2}\eta_{\min}\,||\boldsymbol{u}||^2\,\{1+o_p(1)\} + \frac{b_m}{1+b_m}\sqrt{K\cdot t}\,||\boldsymbol{u}|| + o(1)\right] < 0\right\} \\ ={}& \Pr\left\{\sup_{||\boldsymbol{u}||=M_\varepsilon}\left[\frac{1}{1+b_m}O_p(1)\,||\boldsymbol{u}|| + \frac{b_m}{1+b_m}\sqrt{K\cdot t}\,||\boldsymbol{u}|| + o(1)\right] < \frac{1}{2}\eta_{\min}\,||\boldsymbol{u}||^2\,\{1+o_p(1)\}\right\} \\ ={}& \Pr\left\{\sup_{M_\varepsilon}\left[\frac{1}{1+b_m}O_p(1) + \frac{b_m}{1+b_m}\sqrt{K\cdot t} + o(1)\right] < \frac{1}{2}\eta_{\min}M_\varepsilon\{1+o_p(1)\}\right\} \\ ={}& \Pr\left\{\sup_{M_\varepsilon} O_p(1) < \frac{1}{2}\eta_{\min}M_\varepsilon\{1+o_p(1)\}\right\} \\ \geq{}& 1-\varepsilon, \qquad \text{for sufficiently large } M_\varepsilon \text{ and } m.\end{aligned}$$

Therefore, for any given $\varepsilon > 0$, there exists a sufficiently large M_ε such that

$$\lim_{m\to\infty} \Pr\left\{\sup_{||\boldsymbol{u}||=M_\varepsilon} pl_m(\boldsymbol{\Phi}_0 + r_m\boldsymbol{u}) - pl_m(\boldsymbol{\Phi}_0) < 0\right\} \geq 1-\varepsilon.$$

This completes the proof.

□

Theorem 1 states that when b_m is $O(1)$, there exists a local maximizer $\hat{\boldsymbol{\Phi}}_m$ of the penalized likelihood function (3.3) which has a root-m convergence rate to $\boldsymbol{\Phi}_0$. This can be achieved by the LASSO, Adaptive LASSO and SCAD penalties with proper choice of tuning parameters. For example, if we choose $\lambda_{mk} = O(m^{-1/2})$ for the LASSO and Adaptive LASSO penalties, and $\lambda_{mk} \to 0$ for the SCAD penalty, then it can be seen that $b_m = O(1)$ for all three penalties, and root-m convergence can therefore be achieved.

Theorem 2 below proves that under mild conditions, the penalized likelihood estimators possess the sparsity property which enables consistent variable selection, and are asymptotically normally distributed.

Theorem 2. *Let the observations follow the FMLME model* (3.1) *satisfying regularity conditions A_1 – A_3. Assume that the penalty function p_{mk} satisfies P_0, P_1 and P_2, and that K is known a priori. We have*

(a) For any $\mathbf{\Phi}$ such that $||\mathbf{\Phi} - \mathbf{\Phi}_0|| = O(m^{-1/2})$, with probability tending to 1,

$$pl_m\{(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\} < pl_m\{(\mathbf{\Phi}_1, 0)\}.$$

(b) For any $\sqrt{m}$-consistent maximum penalized likelihood estimator $\hat{\mathbf{\Phi}}_m$ of $\mathbf{\Phi}$,

(i) Sparsity: $\Pr\{\hat{\mathbf{\Phi}}_2 = 0\} \to 1$*, as $m \to \infty$.*

(ii) Asymptotic normality:

$$\sqrt{m}\left[\left\{\boldsymbol{I}_1(\mathbf{\Phi}_{10}) + \frac{\boldsymbol{p}''_m(\mathbf{\Phi}_{10})}{m}\right\}(\hat{\mathbf{\Phi}}_1 - \mathbf{\Phi}_{10}) + \frac{\boldsymbol{p}'_m(\mathbf{\Phi}_{10})}{m}\right] \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_1(\mathbf{\Phi}_{10})),$$

where $I_1(\mathbf{\Phi}_{10})$ is the Fisher information knowing that $\mathbf{\Phi}_2 = 0$.

Proof. (a). Partition $\mathbf{\Phi} = (\mathbf{\Phi}_1, \mathbf{\Phi}_2)$ for any $\mathbf{\Phi}$ in the neighborhood $||\mathbf{\Phi} - \mathbf{\Phi}_0|| = O(m^{-1/2})$. By the definition of $pl_m(\mathbf{\Phi})$, we have

$$\begin{aligned} & pl_m\{(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\} - pl_m\{(\mathbf{\Phi}_1, 0)\} \\ = & [l_m\{(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\} - l_m\{(\mathbf{\Phi}_1, 0)\}] - [\boldsymbol{p}_m\{(\mathbf{\Phi}_1, \mathbf{\Phi}_2)\} - \boldsymbol{p}_m\{(\mathbf{\Phi}_1, 0)\}]. \end{aligned}$$

We now find the order of these two differences. By the mean value theorem,

$$l_m\{(\mathbf{\Phi}_1,\mathbf{\Phi}_2)\} - l_m\{(\mathbf{\Phi}_1,0)\} = \left[\frac{\partial l_m\{(\mathbf{\Phi}_1,\xi)\}}{\partial \mathbf{\Phi}_2}\right]^T \mathbf{\Phi}_2, \tag{5.10}$$

for some $||\xi|| \leq ||\mathbf{\Phi}_2|| = O(m^{-1/2})$. Then,

$$\begin{aligned}\left|\left|\frac{\partial l_m\{(\mathbf{\Phi}_1,\xi)\}}{\partial \mathbf{\Phi}_2} - \frac{\partial l_m\{(\mathbf{\Phi}_{10},0)\}}{\mathbf{\Phi}_2}\right|\right| \leq & \left|\left|\frac{\partial l_m\{(\mathbf{\Phi}_1,\xi)\}}{\partial \mathbf{\Phi}_2} - \frac{\partial l_m\{(\mathbf{\Phi}_1,0)\}}{\mathbf{\Phi}_2}\right|\right| \\ & + \left|\left|\frac{\partial l_m\{(\mathbf{\Phi}_1,0)\}}{\mathbf{\Phi}_2} - \frac{\partial l_m\{(\mathbf{\Phi}_{10},0)\}}{\mathbf{\Phi}_2}\right|\right|. \end{aligned} \tag{5.11}$$

But by the mean value theorem,

$$\begin{aligned} & \frac{\partial l_m\{(\mathbf{\Phi}_1,\xi)\}}{\partial \mathbf{\Phi}_2} - \frac{\partial l_m\{(\mathbf{\Phi}_1,0)\}}{\mathbf{\Phi}_2} = \left[\frac{\partial^2 l_m\{(\mathbf{\Phi}_1,\zeta_1)\}}{\partial \mathbf{\Phi}_2^2}\right] \cdot \xi, \qquad \text{for some } ||\zeta_1|| \leq ||\xi||, \\ \text{and} \quad & \frac{\partial l_m\{(\mathbf{\Phi}_1,0)\}}{\mathbf{\Phi}_2} - \frac{\partial l_m\{(\mathbf{\Phi}_{10},0)\}}{\mathbf{\Phi}_2} = \left[\frac{\partial^2 l_m\{(\zeta_2,0)\}}{\partial \mathbf{\Phi}_1 \mathbf{\Phi}_2}\right] \cdot (\mathbf{\Phi}_1 - \mathbf{\Phi}_0), \end{aligned}$$

where $\zeta_2 = \mathbf{\Phi}_{10} + t \cdot (\mathbf{\Phi}_1 - \mathbf{\Phi}_{10})$, for some $t \in [0,1]$.

Applying these results to (5.11) and using regularity condition A_2, we have

$$\begin{aligned}\left|\left|\frac{\partial l_m\{(\mathbf{\Phi}_1,\xi)\}}{\partial \mathbf{\Phi}_2} - \frac{\partial l_m\{(\mathbf{\Phi}_{10},0)\}}{\mathbf{\Phi}_2}\right|\right| \leq & \left[\sum_{i=1}^{m} M_{2i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)\right] \cdot ||\xi|| \\ & + \left[\sum_{i=1}^{m} M_{2i}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)\right] \cdot ||\mathbf{\Phi}_1 - \mathbf{\Phi}_{10}|| \\ = & \; O_p(m) \cdot \Big(||\xi|| + ||\mathbf{\Phi}_1 - \mathbf{\Phi}_{10}|| \Big) \\ = & \; O_p(m) \cdot \Big\{ O(m^{-\frac{1}{2}}) + O(m^{-\frac{1}{2}}) \Big\} \\ = & \; O_p(m^{\frac{1}{2}}). \end{aligned}$$

By the regularity conditions, $\dfrac{\partial l_m\{(\mathbf{\Phi}_{10},0)\}}{\partial \mathbf{\Phi}_2} = O_p(m^{\frac{1}{2}})$, therefore $\dfrac{\partial l_m\{(\mathbf{\Phi}_1,\xi)\}}{\partial \mathbf{\Phi}_2} = O_p(m^{\frac{1}{2}})$. Using this result on (5.10), we get

$$l_m\{(\mathbf{\Phi}_1,\mathbf{\Phi}_2)\} - l_m\{(\mathbf{\Phi}_1,0)\} = O_p(\sqrt{m}) \sum_{k=1}^{K} \left(\sum_{j=t_{\boldsymbol{\beta}_k}+1}^{p} |\beta_{kj}| + \sum_{j=t_{\boldsymbol{d}_k}+1}^{q} |d_{kj}| \right),$$

where $t_{\boldsymbol{\beta}_k}$ and $t_{\boldsymbol{d}_k}$ are the numbers of true non-zero fixed and random effects in component k, respectively. On the other hand,

$$\boldsymbol{p}_m\{(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\} - \boldsymbol{p}_m\{(\boldsymbol{\Phi}_1, 0)\} = \sum_{k=1}^{K} \left(\sum_{j=t_{\boldsymbol{\beta}_k}+1}^{p} \pi_k \cdot m \cdot p_{mk}(\beta_{kj}) + \sum_{j=t_{\boldsymbol{d}_k}+1}^{q} \pi_k \cdot m \cdot p_{mk}(d_{kj}) \right).$$

Therefore,

$$\begin{aligned} pl_m\{(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\} - pl_m\{(\boldsymbol{\Phi}_1, 0)\} &= \sum_{k=1}^{K} \left[\sum_{j=t_{\boldsymbol{\beta}_k}+1}^{p} \{|\beta_{kj}| \cdot O_p(\sqrt{m}) - \pi_k \cdot m \cdot p_{mk}(\beta_{kj})\} + \right. \\ &\qquad \left. \sum_{j=t_{\boldsymbol{d}_k}+1}^{q} \{|d_{kj}| \cdot O_p(\sqrt{m}) - \pi_k \cdot m \cdot p_{mk}(d_{kj})\} \right] \\ &= \sum_{k=1}^{K} \left[\sum_{j=t_{\boldsymbol{\beta}_k}+1}^{p} A_{kj} + \sum_{j=t_{\boldsymbol{d}_k}+1}^{q} B_{kj} \right], \end{aligned}$$

say. By condition P_2, both A_{kj} and B_{kj} are less than 0 in probability. Therefore,

$$\Pr\left[pl_m\{(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\} - pl_m\{(\boldsymbol{\Phi}_1, 0)\} < 0\right] \xrightarrow{p} 1.$$

This completes the proof of (a).

(b). (i). Let $(\hat{\boldsymbol{\Phi}}_1, 0)$ be the maximizer of the penalized log-likelihood function $pl_m\{(\boldsymbol{\Phi}_1, 0)\}$ which is regarded as a function of $\boldsymbol{\Phi}_1$. It suffices to show that in the neighborhood $||\boldsymbol{\Phi} - \boldsymbol{\Phi}_0|| = O(m^{-1/2})$, the difference $pl_m\{(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\} - pl_m\{(\hat{\boldsymbol{\Phi}}_1, 0)\} < 0$ with probability tending to 1 as $m \to \infty$. We have that

$$\begin{aligned} pl_m\{(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\} - pl_m\{(\hat{\boldsymbol{\Phi}}_1, 0)\} &= [pl_m\{(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\} - pl_m\{(\boldsymbol{\Phi}_1, 0)\}] + [pl_m\{(\boldsymbol{\Phi}_1, 0)\} - pl_m\{(\hat{\boldsymbol{\Phi}}_1, 0)\}] \\ &\le pl_m\{(\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2)\} - pl_m\{(\boldsymbol{\Phi}_1, 0)\} \\ &< 0, \end{aligned}$$

with probability tending to 1 by (a). This completes the proof of (b). (i).

(b). (ii). Regard $pl_m\{(\mathbf{\Phi}_1, 0)\}$ as a function of $\mathbf{\Phi}_1$. Using the same argument as in Theorem 1, there exists a $\sqrt{m}$-consistent local maximizer of this function, say $\hat{\mathbf{\Phi}}_1$, which satisfies

$$\frac{\partial pl_m(\hat{\mathbf{\Phi}}_m)}{\partial \mathbf{\Phi}_1} = \left\{ \frac{\partial l_m(\mathbf{\Phi})}{\partial \mathbf{\Phi}_1} - \frac{\partial \boldsymbol{p}_m(\mathbf{\Phi})}{\partial \mathbf{\Phi}_1} \right\}_{\hat{\mathbf{\Phi}}_m = (\hat{\mathbf{\Phi}}_1, 0)} = 0. \tag{5.12}$$

Since $\hat{\mathbf{\Phi}}_1$ is a $\sqrt{m}$-consistent estimator, by Taylor's expansion around the true value, we have

$$\begin{aligned} \left.\frac{\partial l_m(\mathbf{\Phi})}{\partial \mathbf{\Phi}_1}\right|_{\hat{\mathbf{\Phi}}_m = (\hat{\mathbf{\Phi}}_1, 0)} &= \frac{\partial l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1} + \left\{ \frac{\partial^2 l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1 \partial \mathbf{\Phi}_1^T} + o_p(m) \right\} (\hat{\mathbf{\Phi}}_1 - \mathbf{\Phi}_{10}), \\ \left.\frac{\partial \boldsymbol{p}_m(\mathbf{\Phi})}{\partial \mathbf{\Phi}_1}\right|_{\hat{\mathbf{\Phi}}_m = (\hat{\mathbf{\Phi}}_1, 0)} &= \boldsymbol{p}_m'(\mathbf{\Phi}_{10}) + \{\boldsymbol{p}_m''(\mathbf{\Phi}_{10}) + o_p(m)\}(\hat{\mathbf{\Phi}}_1 - \mathbf{\Phi}_{10}). \end{aligned}$$

Substituting into (5.12), we find

$$\left\{ \frac{\partial l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1} - \boldsymbol{p}_m'(\mathbf{\Phi}_{10}) \right\} + \left\{ \frac{\partial^2 l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1 \partial \mathbf{\Phi}_1^T} - \boldsymbol{p}_m''(\mathbf{\Phi}_{10}) + o_p(m) \right\} (\hat{\mathbf{\Phi}}_1 - \mathbf{\Phi}_{10}) = 0.$$

By rearranging the terms and multiplying both sides by $1/\sqrt{m}$, we get

$$\sqrt{m} \cdot -\frac{1}{m} \left\{ \frac{\partial^2 l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1 \partial \mathbf{\Phi}_1^T} - \boldsymbol{p}_m''(\mathbf{\Phi}_{10}) + o_p(m) \right\} (\hat{\mathbf{\Phi}}_1 - \mathbf{\Phi}_{10}) = \frac{1}{\sqrt{m}} \left\{ \frac{\partial l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1} - \boldsymbol{p}_m'(\mathbf{\Phi}_{10}) \right\}.$$

Then, by the regularity conditions,

$$-\frac{1}{m} \frac{\partial^2 l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1 \partial \mathbf{\Phi}_1^T} = I_1(\mathbf{\Phi}_{10}) + o_p(1), \qquad \frac{1}{\sqrt{m}} \frac{\partial l_m(\mathbf{\Phi}_{10})}{\partial \mathbf{\Phi}_1} \xrightarrow{d} \mathcal{N}(0, I_1(\mathbf{\Phi}_{10})).$$

Thus by Slutsky's theorem,

$$\sqrt{m} \left[\left\{ I_1(\mathbf{\Phi}_{10}) + \frac{\boldsymbol{p}_m''(\mathbf{\Phi}_{10})}{m} \right\} (\hat{\mathbf{\Phi}}_1 - \mathbf{\Phi}_{10}) + \frac{\boldsymbol{p}_m'(\mathbf{\Phi}_{10})}{m} \right] \xrightarrow{d} \mathcal{N}(0, I_1(\mathbf{\Phi}_{10})).$$

□

For the Adaptive LASSO and SCAD penalties, sparsity could be achieved while maintaining root-m consistency, for suitable choice of tuning parameters. For example, if we let $\lambda_{mk} = O(m^{-1/2})$ for the Adaptive LASSO penalty, and $\lambda_{mk} \to 0$ and $\sqrt{m}\lambda_{mk} \to \infty$ for the SCAD penalty, then root-m consistency and sparsity can be achieved concurrently. This is, however, not true for the LASSO penalty. For the latter, $b_m = \sqrt{m}\lambda_{mk}$. Therefore, the root-$m$ consistency requires that $\sqrt{m}\lambda_{mk} = O(1)$. On the other hand, the sparsity property requires assumption P_2, which includes the condition $\sqrt{m}\lambda_{mk} \to \infty$. These two requirements cannot be simultaneously satisfied.

The derivatives of $\boldsymbol{p}_m$ in part b(ii) of Theorem 2 become negligible for some choices of the penalty function, thus the result suggests the following variance estimator of $\hat{\boldsymbol{\Phi}}_1$:

$$\hat{\text{Var}}\,(\hat{\boldsymbol{\Phi}}_1) = \left\{ l''_m(\hat{\boldsymbol{\Phi}}_1) - p''_m(\hat{\boldsymbol{\Phi}}_1) \right\}^{-1} \hat{\text{Var}}\, \left\{ l'_m(\hat{\boldsymbol{\Phi}}_1) \right\} \left\{ l''_m(\hat{\boldsymbol{\Phi}}_1) - p''_m(\hat{\boldsymbol{\Phi}}_1) \right\}^{-1}.$$

However, due to the super-efficiency phenomenon associated with model selection, the conclusions on asymptotic bias or variance should be used cautiously (Leeb & Pötscher, 2003).

Having derived the asymptotic properties of the penalized estimator, we examine in the next chapter the finite-sample performance of our method through simulations.

CHAPTER 6
Simulation study

We perform simulations to investigate the finite sample performance of our method. We consider the two-component FMLME model

$$\pi\mathcal{N}\big(\boldsymbol{X}_i\boldsymbol{\beta}_1, \sigma_1^2(\boldsymbol{Z}_i\boldsymbol{\Psi}_1\boldsymbol{Z}_i' + I_{n_i})\big) + (1-\pi)\mathcal{N}\big(\boldsymbol{X}_i\boldsymbol{\beta}_2, \sigma_2^2(\boldsymbol{Z}_i\boldsymbol{\Psi}_2\boldsymbol{Z}_i' + I_{n_i})\big),$$

with two different parameter settings. For the first simulations, the parameters are as follows:

– **Model 1**:

$$\boldsymbol{\beta}_1 = (2.5, 2, 0, 0, 0, 0, 0, 0, 0) = (\beta_{1_1}, \beta_{1_2}, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\beta}_2 = (-1, -1.5, 0, 0, 0, 0, 0, 0, 0) = (\beta_{2_1}, \beta_{2_2}, 0, 0, 0, 0, 0, 0, 0),$$

$$\boldsymbol{b}_{1i} = (b_{1i_1}, b_{1i_2}, 0, 0), \quad \boldsymbol{b}_{2i} = (b_{2i_1}, b_{2i_2}, 0, 0), \qquad \text{and} \qquad \sigma_1^2 = 1.5, \quad \sigma_2^2 = 0.5,$$

where the covariance matrices of the truly significant random effects of the two components are given by

$$\sigma_1^2 \cdot \begin{bmatrix} \psi_{1_{11}} & \psi_{1_{12}} \\ \psi_{1_{12}} & \psi_{1_{22}} \end{bmatrix} = \sigma_1^2 \cdot \begin{bmatrix} 9 & 4.8 \\ 4.8 & 4 \end{bmatrix} \qquad \text{and} \qquad \sigma_2^2 \cdot \begin{bmatrix} \psi_{2_{11}} & \psi_{2_{12}} \\ \psi_{2_{12}} & \psi_{2_{22}} \end{bmatrix} = \sigma_2^2 \cdot \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix}.$$

– **Model 2**: we consider the same variance structures as in Model 1, but we changed both component-wise means. This allows us to investigate the impact of significant random effects on the estimation and variable selection of insignificant fixed effects. The parameters are as follows.

$$\boldsymbol{\beta}_1 = (2.5, 0, 0, 2, 0, 0, 0, 0, 0) = (\beta_{1_1}, 0, 0, \beta_{1_2}, 0, 0, 0, 0, 0),$$

$$\boldsymbol{\beta}_2 = (0, -1, 0, 0, 0, 0, 0, 0, -1.5) = (0, \beta_{2_1}, 0, 0, 0, 0, 0, 0, \beta_{2_2}),$$

$\boldsymbol{b}_{1i} = (b_{1i_1}, b_{1i_2}, 0, 0), \quad \boldsymbol{b}_{2i} = (b_{2i_1}, b_{2i_2}, 0, 0),$

the covariance matrices of the random effects and errors are the same as in Model 1.

For both models, the design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ for each subject are generated separately from multivariate normal distributions with mean 0 and an AR type variance matrix with $\rho_{ij} = \text{Corr}\,(x_i, x_j) = \text{Corr}\,(z_i, z_j)$, and are subsequently standardized. The first column of $\boldsymbol{Z}_i$ consists of a column of 1's.

For each of the above two models, we consider 3 sample sizes: $(m = 80,\ n_i = 5)$, $(m = 120,\ n_i = 5)$ and $(m = 200,\ n_i = 5)$. For each of these sample sizes, we consider both balanced mixtures ($\pi_1 = 0.5$) and unbalanced mixtures ($\pi_1 = 0.3$). Finally, for each of these mixing proportion settings, we consider a moderate correlation case with $\rho_{ij} = 0.5^{|i-j|}$ and a high correlation case with $\rho_{ij} = 0.75^{|i-j|}$. This creates a total of 24 scenarios, which allows us to examine the method's performance across a wide scope of situations. The simulation results, provided at the end of the chapter from page 46 to 55, are based on 200 data sets for each scenario.

To examine the variable selection performance of the method, we consider the sensitivity and specificity of selection. The former is defined as the conditional probability that an effect is selected given that it is truly significant, whereas the latter is the conditional probability that an effect is removed given that it is truly non-significant. For both criteria, a higher value indicates better selection performance. It turns out that the empirical sensitivity for all simulation scenarios is 100%, which means that the important effects are always kept in the model using our method. The empirical specificity is presented in Tables 6–1 to 6–4. To examine the estimation accuracy, we report the the empirical mean squared errors (MSE) of the penalized estimates of non-zero parameters, and the empirical MSE of the Oracle estimates of non-zero parameters, which are ML estimates of the parameters knowing the true model in advance (Tables 6–5 to 6–12). Let $\hat{\theta}_i$ be the estimate of a truly non-zero parameter θ based on the i-th data set, then the empirical MSE based on n simulated data sets is found by

$$MSE_n = \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2,$$

where $n = 200$ in all of our simulation scenarios.

From Table 6–1 to Table 6–4, we observe that the Adaptive LASSO is the best of the three penalties in eliminating the truly zero effects, followed by SCAD, whereas LASSO is the least effective. We also observe that, for any given correlation ρ_{ij} in the design matrices and mixing proportions, the empirical specificity improves as the number of subjects m increases. For example, it can be seen in Table 6–1 that, when m reaches 200 for the balanced mixture, the empirical specificity of the Adaptive LASSO penalty is 97.6% for the first component and 100% for the second component. Then, for any given sample size and correlation ρ_{ij}, when π_1 decreases, all three penalties perform less satisfactorily in the first component due to the lower number of observations. For all scenarios, all three penalties perform better in the second component than the first. This is presumably because the random effects of the latter have greater variances, which renders variable selection more difficult. Fixing the mixing proportions, larger correlation ρ_{ij} also makes the selection harder. Finally, given the same correlation, mixing proportions and sample size, the three penalties perform similarly in Model 1 and Model 2.

From Tables 6–5 to 6–12, we observe that the estimates obtained using the Adaptive LASSO penalty are the closest to the Oracle model in terms of MSE. We also note that occasionally the Adaptive LASSO estimates have a slightly better MSE than the oracle estimates, which is most likely due to numerical errors. The LASSO and SCAD penalties have similar estimation accuracy for the fixed effects, but for the random effect variances, SCAD performs less satisfactorily than LASSO, especially for ψ_{11} in the first component. We also observe that the estimation improves as the number of subjects m increases, which confirms the consistency property of the estimators. By looking at the performance when

$m = 80$, we see that the penalized estimators for the fixed effect parameters β's achieve a fairly good estimation accuracy, but the penalized estimators for the variance components are less reliable. Furthermore, the fixed effects are selected correctly less often when $m = 80$. These observations suggest that in applications where the sample size is small, one should interpret the results with care. Then, it can be seen that MSE's increase when π_1 decreases, indicating that estimation becomes more difficult in the first component due to the lower number of observations. For all scenarios, we also observe that MSE's are much smaller for the fixed effects than for the variance components. The larger random effect variances of the first component also seem to result in larger MSE's. We can also see that fixing the mixing proportions, larger correlation ρ_{ij} makes the estimation less accurate. Finally, for the same correlation, mixing proportions and sample size, estimation is slightly better in Model 1 than in Model 2, but the difference is not substantial.

Table 6–1: Empirical specificity of selection based on 200 random samples from Model 1 of sizes 80×5, 120×5 and 200×5, with $\rho_{ij} = 0.5$.

m	Component	Effect	Empirical Specificity (in %)			Empirical Specificity (in %)		
			$(\pi_1, \pi_2) = (0.3, 0.7)$			$(\pi_1, \pi_2) = (0.5, 0.5)$		
			LASSO	ALASSO	SCAD	LASSO	ALASSO	SCAD
80	1	Fixed	48.4	73.5	50.0	60.4	84.4	62.6
		Random	96.8	100	98.8	98.8	100	98.8
	2	Fixed	91.1	99.4	93.6	85.9	96.5	87.5
		Random	98.8	100	99.8	99.2	100	100
120	1	Fixed	57.5	81.8	59.2	71.7	92.6	75.3
		Random	98.0	100	98.2	98.0	100	99.2
	2	Fixed	96.0	99.5	97.1	92.1	98.9	93.1
		Random	99.5	100	99.5	99.2	100	100
200	1	Fixed	71.6	91.1	73.1	85.3	97.6	86.7
		Random	99.5	100	98.5	99.8	100	99.2
	2	Fixed	99.2	99.9	99.2	98.1	100	98.1
		Random	99.2	100	99.8	100	100	100

Table 6–2: Empirical specificity of selection based on 200 random samples from Model 1 of sizes 80×5, 120×5 and 200×5, with $\rho_{ij} = 0.75$.

m	Component	Effect	Empirical Specificity (in %)			Empirical Specificity (in %)		
			$(\pi_1, \pi_2) = (0.3, 0.7)$			$(\pi_1, \pi_2) = (0.5, 0.5)$		
			LASSO	ALASSO	SCAD	LASSO	ALASSO	SCAD
80	1	Fixed	41.9	65.2	44.5	56.1	79.8	62.8
		Random	96.2	100	98.2	97.0	100	97.8
	2	Fixed	83.0	98.4	88.1	79.5	96.8	85.1
		Random	99.5	100	99.8	98.5	100	99.8
120	1	Fixed	48.9	74.6	51.0	56.1	79.8	62.8
		Random	97.0	99.8	97.8	99.2	100	99.2
	2	Fixed	92.1	99.6	94.9	88.1	98.4	92.7
		Random	100	100	99.5	100	100	99.8
200	1	Fixed	65.9	87.3	68.7	76.4	96.5	86.0
		Random	98.0	100	98.0	98.8	100	99.5
	2	Fixed	97.7	99.9	98.7	96.5	100	97.4
		Random	100	100	100	99.8	100	100

Table 6–3: Empirical specificity of selection based on 200 random samples from Model 2 of sizes 80×5, 120×5 and 200×5, with $\rho_{ij} = 0.5$.

m	Component	Effect	Empirical Specificity (in %)			Empirical Specificity (in %)		
			$(\pi_1, \pi_2) = (0.3, 0.7)$			$(\pi_1, \pi_2) = (0.5, 0.5)$		
			LASSO	ALASSO	SCAD	LASSO	ALASSO	SCAD
80	1	Fixed	48.8	72.6	52.2	57.2	83.6	63.9
		Random	97.5	100	99.0	99.2	99.8	99.2
	2	Fixed	89.5	98.3	91.9	84.0	98.0	85.9
		Random	98.0	100	100	99.5	100	100
120	1	Fixed	60.3	81.2	61.4	72.6	92.1	76.9
		Random	99.0	100	99.8	99.0	100	99.2
	2	Fixed	96.6	99.6	97.9	90.1	99.1	92.0
		Random	98.8	100	100	98.8	100	99.8
200	1	Fixed	72.1	91.1	72.6	84.4	96.7	88.1
		Random	99.5	100	99.2	99.8	100	99.8
	2	Fixed	99.1	99.9	99.1	97.6	99.6	98.5
		Random	99.0	100	99.8	99.5	100	100

Table 6–4: Empirical specificity of selection based on 200 random samples from Model 2 of sizes 80×5, 120×5 and 200×5, with $\rho_{ij} = 0.75$.

m	Component	Effect	Empirical Specificity (in %)			Empirical Specificity (in %)		
			$(\pi_1, \pi_2) = (0.3, 0.7)$			$(\pi_1, \pi_2) = (0.5, 0.5)$		
			LASSO	ALASSO	SCAD	LASSO	ALASSO	SCAD
80	1	Fixed	44.7	68.2	46.5	51.2	80.1	61.4
		Random	96.2	100	97.8	98.2	100	99.2
	2	Fixed	83.2	98.4	91.5	77.1	95.9	84.8
		Random	99.0	100	99.5	99.0	100	100
120	1	Fixed	48.3	76.2	55.0	67.4	90.6	78.6
		Random	98.8	100	98.5	99.2	100	99.8
	2	Fixed	90.6	99.5	94.1	85.7	98.3	89.7
		Random	100	100	100	99.8	100	100
200	1	Fixed	64.7	89.1	69.7	78.2	95.7	85.4
		Random	99.0	100	99.2	99.5	100	99.8
	2	Fixed	96.4	99.9	98.4	94.1	99.6	97.1
		Random	99.8	100	100	100	100	100

Table 6–5: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 1, with $\rho_{ij} = 0.5$, and $(\pi_1, \pi_2) = (0.5, 0.5)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.017	0.017	5.063	1.811	1.216	0.042	0.004
		ALASSO	0.016	0.015	5.713	2.017	1.193	0.035	0.004
		SCAD	0.017	0.017	9.342	3.143	2.028	0.053	0.004
		ORACLE	0.016	0.014	5.847	2.056	1.190	0.032	0.004
	2	LASSO	0.005	0.007	3.026	0.701	0.840	0.004	0.004
		ALASSO	0.005	0.006	2.566	0.710	0.755	0.004	0.004
		SCAD	0.006	0.007	3.992	1.032	1.141	0.005	0.004
		ORACLE	0.005	0.006	3.206	0.888	0.924	0.004	0.004
120	1	LASSO	0.011	0.011	3.092	1.145	0.858	0.024	0.002
		ALASSO	0.010	0.009	3.378	1.169	0.723	0.022	0.002
		SCAD	0.011	0.010	4.789	1.704	1.134	0.030	0.002
		ORACLE	0.010	0.008	3.748	1.309	0.776	0.023	0.002
	2	LASSO	0.004	0.004	2.410	0.452	0.591	0.004	0.002
		ALASSO	0.004	0.003	1.754	0.424	0.480	0.003	0.002
		SCAD	0.004	0.004	2.017	0.514	0.577	0.003	0.002
		ORACLE	0.004	0.003	1.778	0.467	0.503	0.003	0.002
200	1	LASSO	0.007	0.009	2.053	0.753	0.533	0.016	0.001
		ALASSO	0.007	0.006	2.034	0.753	0.525	0.016	0.001
		SCAD	0.007	0.008	2.518	0.913	0.681	0.020	0.001
		ORACLE	0.007	0.006	2.089	0.769	0.543	0.016	0.001
	2	LASSO	0.002	0.003	2.067	0.344	0.518	0.003	0.001
		ALASSO	0.002	0.002	1.125	0.294	0.325	0.002	0.001
		SCAD	0.002	0.002	1.132	0.327	0.307	0.002	0.001
		ORACLE	0.002	0.002	1.024	0.301	0.275	0.002	0.001

Table 6–6: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 1, with $\rho_{ij} = 0.75$, and $(\pi_1, \pi_2) = (0.5, 0.5)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.029	0.042	5.244	2.203	1.443	0.046	0.003
		ALASSO	0.027	0.032	5.368	2.097	1.243	0.034	0.003
		SCAD	0.028	0.038	8.289	3.294	2.045	0.053	0.003
		ORACLE	0.026	0.022	5.297	2.045	1.227	0.031	0.003
	2	LASSO	0.007	0.009	2.817	0.547	0.700	0.004	0.003
		ALASSO	0.007	0.007	2.377	0.627	0.698	0.004	0.003
		SCAD	0.007	0.009	3.007	0.823	0.966	0.005	0.003
		ORACLE	0.007	0.006	2.515	0.740	0.819	0.004	0.003
120	1	LASSO	0.020	0.026	3.858	1.249	0.803	0.031	0.002
		ALASSO	0.019	0.023	4.153	1.401	0.863	0.027	0.002
		SCAD	0.019	0.025	5.201	1.735	1.116	0.035	0.002
		ORACLE	0.018	0.021	4.375	1.512	0.929	0.028	0.002
	2	LASSO	0.006	0.009	2.925	0.497	0.674	0.005	0.002
		ALASSO	0.006	0.007	2.060	0.451	0.544	0.004	0.002
		SCAD	0.006	0.008	2.172	0.498	0.568	0.003	0.002
		ORACLE	0.006	0.006	1.970	0.474	0.514	0.003	0.002
200	1	LASSO	0.012	0.016	2.436	0.905	0.579	0.013	0.001
		ALASSO	0.012	0.012	2.311	0.844	0.539	0.014	0.001
		SCAD	0.011	0.014	2.653	1.034	0.641	0.016	0.001
		ORACLE	0.011	0.011	2.412	0.893	0.547	0.014	0.001
	2	LASSO	0.003	0.006	2.140	0.325	0.538	0.003	0.001
		ALASSO	0.003	0.004	1.068	0.275	0.363	0.002	0.001
		SCAD	0.003	0.005	1.212	0.329	0.400	0.002	0.001
		ORACLE	0.003	0.003	1.120	0.319	0.384	0.002	0.001

Table 6–7: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 1, with $\rho_{ij} = 0.5$, and $(\pi_1, \pi_2) = (0.3, 0.7)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.031	0.039	11.998	4.182	3.022	0.107	0.002
		ALASSO	0.029	0.037	14.400	4.568	2.981	0.082	0.002
		SCAD	0.030	0.039	25.412	7.778	5.243	0.131	0.002
		ORACLE	0.028	0.033	12.565	3.959	2.516	0.067	0.002
	2	LASSO	0.004	0.005	1.965	0.551	0.690	0.003	0.002
		ALASSO	0.003	0.004	1.562	0.490	0.619	0.003	0.002
		SCAD	0.004	0.005	2.689	0.694	0.853	0.004	0.002
		ORACLE	0.003	0.004	2.109	0.591	0.721	0.003	0.002
120	1	LASSO	0.017	0.023	7.421	2.918	2.197	0.064	0.002
		ALASSO	0.016	0.020	9.360	3.387	2.289	0.057	0.002
		SCAD	0.017	0.023	15.187	5.360	3.772	0.081	0.002
		ORACLE	0.016	0.017	9.379	3.387	2.291	0.054	0.002
	2	LASSO	0.003	0.003	1.698	0.334	0.489	0.003	0.002
		ALASSO	0.003	0.003	1.149	0.291	0.376	0.002	0.002
		SCAD	0.003	0.003	1.420	0.354	0.459	0.002	0.002
		ORACLE	0.003	0.003	1.206	0.323	0.387	0.002	0.002
200	1	LASSO	0.014	0.016	3.293	1.209	0.846	0.030	0.001
		ALASSO	0.014	0.013	3.427	1.199	0.781	0.029	0.001
		SCAD	0.014	0.015	4.919	1.742	1.113	0.037	0.001
		ORACLE	0.013	0.011	3.582	1.258	0.798	0.029	0.001
	2	LASSO	0.002	0.002	1.347	0.237	0.354	0.002	0.001
		ALASSO	0.002	0.002	0.744	0.191	0.230	0.002	0.001
		SCAD	0.002	0.002	0.819	0.211	0.263	0.001	0.001
		ORACLE	0.002	0.002	0.778	0.204	0.236	0.001	0.001

Table 6–8: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 1, with $\rho_{ij} = 0.75$, and $(\pi_1, \pi_2) = (0.3, 0.7)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.058	0.088	11.353	3.599	2.697	0.092	0.003
		ALASSO	0.058	0.077	12.674	3.760	2.289	0.075	0.003
		SCAD	0.058	0.086	21.986	6.671	4.640	0.111	0.003
		ORACLE	0.053	0.046	11.024	3.268	1.959	0.063	0.003
	2	LASSO	0.006	0.010	2.191	0.485	0.721	0.004	0.003
		ALASSO	0.007	0.007	1.907	0.496	0.585	0.003	0.003
		SCAD	0.006	0.009	2.528	0.633	0.690	0.004	0.003
		ORACLE	0.006	0.006	2.217	0.570	0.621	0.003	0.003
120	1	LASSO	0.027	0.043	6.373	2.706	2.023	0.048	0.002
		ALASSO	0.025	0.036	6.576	2.744	1.779	0.036	0.002
		SCAD	0.027	0.041	10.516	4.046	2.931	0.059	0.002
		ORACLE	0.024	0.024	6.469	2.668	1.698	0.033	0.002
	2	LASSO	0.004	0.005	1.900	0.367	0.569	0.003	0.002
		ALASSO	0.004	0.004	1.100	0.297	0.423	0.002	0.002
		SCAD	0.003	0.005	1.344	0.351	0.433	0.002	0.002
		ORACLE	0.003	0.004	1.198	0.316	0.403	0.002	0.002
200	1	LASSO	0.020	0.024	3.358	1.134	0.905	0.024	0.001
		ALASSO	0.019	0.019	3.646	1.178	0.784	0.023	0.001
		SCAD	0.019	0.023	4.770	1.659	1.204	0.030	0.001
		ORACLE	0.019	0.015	3.830	1.249	0.801	0.025	0.001
	2	LASSO	0.003	0.004	1.679	0.258	0.470	0.003	0.001
		ALASSO	0.003	0.003	0.825	0.199	0.251	0.002	0.001
		SCAD	0.003	0.003	0.810	0.231	0.260	0.001	0.001
		ORACLE	0.003	0.003	0.773	0.222	0.248	0.001	0.001

Table 6–9: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 2, with $\rho_{ij} = 0.5$, and $(\pi_1, \pi_2) = (0.5, 0.5)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.020	0.023	5.231	1.920	1.369	0.044	0.003
		ALASSO	0.016	0.016	5.754	2.137	1.486	0.047	0.003
		SCAD	0.018	0.021	7.873	2.942	2.092	0.051	0.003
		ORACLE	0.014	0.012	5.819	2.093	1.482	0.051	0.003
	2	LASSO	0.008	0.007	2.833	0.669	0.813	0.005	0.003
		ALASSO	0.005	0.006	2.663	0.706	0.818	0.007	0.003
		SCAD	0.007	0.006	3.212	0.907	0.937	0.005	0.003
		ORACLE	0.004	0.005	2.782	0.770	0.795	0.006	0.003
120	1	LASSO	0.009	0.012	4.247	1.398	0.970	0.030	0.002
		ALASSO	0.008	0.010	4.737	1.518	1.003	0.029	0.002
		SCAD	0.009	0.011	6.214	1.912	1.289	0.034	0.002
		ORACLE	0.008	0.008	4.999	1.590	1.033	0.028	0.002
	2	LASSO	0.004	0.004	2.578	0.428	0.683	0.005	0.002
		ALASSO	0.003	0.003	1.942	0.409	0.565	0.006	0.002
		SCAD	0.004	0.003	2.012	0.473	0.574	0.004	0.002
		ORACLE	0.002	0.003	1.837	0.439	0.493	0.004	0.002
200	1	LASSO	0.007	0.008	2.403	0.722	0.497	0.017	0.001
		ALASSO	0.006	0.006	2.197	0.681	0.475	0.018	0.001
		SCAD	0.006	0.007	2.698	0.905	0.613	0.017	0.001
		ORACLE	0.006	0.005	2.257	0.723	0.490	0.016	0.001
	2	LASSO	0.003	0.002	2.062	0.339	0.519	0.004	0.001
		ALASSO	0.002	0.001	1.192	0.272	0.326	0.003	0.001
		SCAD	0.003	0.002	1.195	0.299	0.311	0.002	0.001
		ORACLE	0.002	0.001	1.093	0.284	0.290	0.002	0.001

Table 6–10: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 2, with $\rho_{ij} = 0.75$, and $(\pi_1, \pi_2) = (0.5, 0.5)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.029	0.037	4.654	1.750	1.343	0.039	0.003
		ALASSO	0.018	0.025	4.841	1.667	1.143	0.031	0.003
		SCAD	0.026	0.034	7.033	2.500	1.874	0.045	0.003
		ORACLE	0.012	0.011	4.992	1.717	1.187	0.032	0.003
	2	LASSO	0.013	0.009	2.753	0.584	0.743	0.004	0.003
		ALASSO	0.006	0.005	2.238	0.593	0.612	0.004	0.003
		SCAD	0.010	0.008	2.962	0.799	0.751	0.004	0.003
		ORACLE	0.004	0.004	2.383	0.683	0.636	0.004	0.003
120	1	LASSO	0.017	0.030	3.350	1.187	0.854	0.023	0.002
		ALASSO	0.012	0.018	3.974	1.418	0.947	0.026	0.002
		SCAD	0.014	0.024	4.878	1.686	1.155	0.030	0.002
		ORACLE	0.009	0.009	4.127	1.484	1.004	0.026	0.002
	2	LASSO	0.008	0.007	2.581	0.475	0.698	0.004	0.002
		ALASSO	0.003	0.004	1.804	0.450	0.551	0.003	0.002
		SCAD	0.006	0.005	1.926	0.537	0.601	0.003	0.002
		ORACLE	0.002	0.003	1.779	0.494	0.536	0.003	0.002
200	1	LASSO	0.012	0.014	2.235	0.720	0.514	0.014	0.001
		ALASSO	0.008	0.008	2.060	0.669	0.465	0.015	0.001
		SCAD	0.010	0.011	2.336	0.747	0.517	0.016	0.001
		ORACLE	0.007	0.006	2.170	0.696	0.466	0.014	0.001
	2	LASSO	0.006	0.003	2.406	0.395	0.613	0.004	0.001
		ALASSO	0.002	0.002	1.230	0.299	0.407	0.003	0.001
		SCAD	0.003	0.002	1.103	0.318	0.365	0.002	0.001
		ORACLE	0.002	0.002	1.053	0.298	0.341	0.002	0.001

Table 6–11: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 2, with $\rho_{ij} = 0.5$, and $(\pi_1, \pi_2) = (0.3, 0.7)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.034	0.042	18.453	5.852	3.574	0.096	0.003
		ALASSO	0.027	0.032	15.886	5.122	3.358	0.080	0.003
		SCAD	0.032	0.039	22.130	7.235	5.016	0.106	0.003
		ORACLE	0.022	0.021	12.905	4.286	2.907	0.082	0.003
	2	LASSO	0.005	0.005	2.540	0.500	0.689	0.004	0.003
		ALASSO	0.003	0.004	1.827	0.457	0.568	0.004	0.003
		SCAD	0.004	0.005	1.989	0.511	0.625	0.003	0.003
		ORACLE	0.003	0.004	1.804	0.475	0.551	0.003	0.003
120	1	LASSO	0.018	0.020	6.148	2.021	1.279	0.049	0.002
		ALASSO	0.015	0.016	7.165	2.292	1.433	0.053	0.002
		SCAD	0.017	0.020	9.540	2.925	1.929	0.061	0.002
		ORACLE	0.014	0.012	6.889	2.199	1.383	0.054	0.002
	2	LASSO	0.003	0.002	1.494	0.367	0.437	0.003	0.002
		ALASSO	0.002	0.001	1.207	0.348	0.409	0.003	0.002
		SCAD	0.003	0.002	1.264	0.405	0.401	0.002	0.002
		ORACLE	0.002	0.001	1.174	0.371	0.366	0.002	0.002
200	1	LASSO	0.011	0.016	3.649	1.290	0.825	0.025	0.001
		ALASSO	0.009	0.012	3.939	1.397	0.850	0.026	0.001
		SCAD	0.010	0.016	5.267	1.789	1.156	0.030	0.001
		ORACLE	0.008	0.009	4.090	1.452	0.881	0.026	0.001
	2	LASSO	0.002	0.002	1.076	0.188	0.326	0.002	0.001
		ALASSO	0.001	0.001	0.703	0.159	0.229	0.001	0.001
		SCAD	0.002	0.002	0.799	0.186	0.242	0.001	0.001
		ORACLE	0.001	0.001	0.702	0.163	0.216	0.001	0.001

Table 6–12: Empirical MSE of parameter estimates based on 200 random samples of sizes 80 × 5, 120 × 5 and 200 × 5 from Model 2, with $\rho_{ij} = 0.75$, and $(\pi_1, \pi_2) = (0.3, 0.7)$.

m	Component	Penalty	MSE of parameter estimates						
			β_1	β_2	ψ_{11}	ψ_{12}	ψ_{22}	σ^2	π
80	1	LASSO	0.049	0.058	10.156	4.472	3.589	0.082	0.002
		ALASSO	0.039	0.051	13.537	5.050	3.490	0.069	0.002
		SCAD	0.048	0.065	22.805	8.111	6.151	0.103	0.002
		ORACLE	0.026	0.024	13.042	4.669	3.123	0.072	0.002
	2	LASSO	0.009	0.006	2.465	0.569	0.625	0.004	0.002
		ALASSO	0.004	0.003	2.161	0.543	0.565	0.004	0.002
		SCAD	0.006	0.005	2.919	0.671	0.668	0.004	0.002
		ORACLE	0.003	0.003	2.554	0.609	0.587	0.004	0.002
120	1	LASSO	0.029	0.038	5.400	1.803	1.379	0.050	0.002
		ALASSO	0.022	0.026	6.608	2.018	1.320	0.042	0.002
		SCAD	0.028	0.037	9.531	2.787	2.067	0.060	0.002
		ORACLE	0.015	0.014	6.738	2.070	1.370	0.041	0.002
	2	LASSO	0.005	0.004	2.147	0.355	0.553	0.003	0.002
		ALASSO	0.002	0.002	1.475	0.288	0.385	0.002	0.002
		SCAD	0.004	0.003	1.604	0.355	0.427	0.002	0.002
		ORACLE	0.002	0.002	1.486	0.322	0.363	0.002	0.002
200	1	LASSO	0.019	0.024	3.475	1.171	0.864	0.024	0.001
		ALASSO	0.015	0.016	4.054	1.307	0.917	0.025	0.001
		SCAD	0.017	0.022	5.634	1.814	1.336	0.030	0.001
		ORACLE	0.013	0.009	4.261	1.373	0.912	0.024	0.001
	2	LASSO	0.004	0.004	1.767	0.278	0.497	0.004	0.001
		ALASSO	0.002	0.002	0.754	0.190	0.243	0.002	0.001
		SCAD	0.002	0.002	0.722	0.214	0.230	0.001	0.001
		ORACLE	0.001	0.001	0.699	0.202	0.210	0.001	0.001

CHAPTER 7
Real data analysis

The object of the SSc study conducted by the Canadian Scleroderma Research Group is to assess disability in SSc patients and its association with various clinical characteristics of interest. For our analysis, in order to obtain more conclusive assessment on the progression of disability, only patients with at least for 4 visits and complete baseline demographic information were included. This yields a total number of 1982 observations from 378 patients. The outcome is measured using HAQ, which ranges from 0 (no disability) to 3 (severe disability). The baseline characteristics of interest include age, gender, disease duration, and baseline HAQ. The time-dependent variables of interest include the severity of skin hardening, breathing problems, gastrointestinal symptoms, Raynaud's phenomenon, digital ulcers, tender joint counts, and visit number. For the precise definitions of these clinical characteristics, we refer the reader to Schnitzer et al. (2011). We fit a two-component FMLME model, where the fixed effects include all the above variables plus an intercept. For the random component, we allow for a random intercept and a possible random slope for all the time-dependent variables. The responses and all continuous predictors are standardized in our analysis.

We first find the ML estimates of our two-component FMLME model. The estimated mixing proportions are $(\hat{\pi}_1, \hat{\pi}_2) = (0.56, 0.44)$, and suggest that we have an almost balanced mixture. The error variances are estimated to be $(\hat{\sigma}_1^2, \hat{\sigma}_2^2) = (0.035, 0.188)$. The fact that $\hat{\sigma}_2^2$ is five times larger than $\hat{\sigma}_1^2$ implies that we might have more variability in the second component. Table 7–1 presents the ML estimates of fixed effects and their standard errors found using the empirical information matrix, and the random effect standard deviations.

From Table 7–1, we see that many of the fixed effect parameter estimates are quite large compared to their standard errors, indicating that these effects are potentially insignificant and that variable selection is required.

Table 7–1: ML estimates for the two-component FMLME model.

Variable	Component 1 ($\hat{\pi}_1 = 0.56$, $\hat{\sigma}_1^2 = 0.035$)		Component 2 ($\hat{\pi}_2 = 0.44$, $\hat{\sigma}_2^2 = 0.188$)	
	Fixed (Std. Error)	Random effect SD	Fixed (Std. Error)	Random effect SD
`Intercept`	-0.059 (0.018)	0.189	0.132 (0.051)	0.457
`Visit`	0.007 (0.013)	0.117	0.164 (0.029)	0.226
`Male`	-0.003 (0.047)	–	-0.228 (0.143)	–
`Baseline Age`	-0.003 (0.015)	–	0.001 (0.044)	–
`Baseline Duration`	0.001 (0.019)	–	-0.047 (0.047)	–
`Baseline HAQ`	0.923 (0.017)	–	0.583 (0.042)	–
`Skin Score`	0.013 (0.014)	0.030	0.090 (0.030)	0.112
`Short Breath`	0.029 (0.016)	0.085	0.132 (0.032)	0.148
`Gastro`	-0.001 (0.012)	0.050	0.038 (0.029)	0.098
`Raynaud`	0.010 (0.012)	0.013	0.138 (0.027)	0.122
`Tender Joints`	0.006 (0.010)	0.015	0.008 (0.017)	0.035
`Ulcers`	0.005 (0.011)	0.006	0.004 (0.033)	0.038

We then fit the two-component FMLME model using the LASSO, Adaptive LASSO and SCAD penalties. The penalized estimates are reported in Tables 7–2, 7–3 and 7–4 respectively. From these tables, we see that the penalized estimates of mixing proportions and error variances are all very close to their corresponding ML estimates, but the three penalized models are much more parsimonious than the ML model in Table 7–1. Furthermore, in all three penalized models, the mean structures and variance components in the two mixture components are quite different, which reiterates the importance of variable selection in each component of an FMLME model. Comparing Tables 7–2, 7–3 and 7–4, we see that the Adaptive LASSO and SCAD penalties choose almost the same fixed and random effects, and produce very similar parameter estimates. On the other hand, we obtain a slightly more complex model using the LASSO penalty. This is presumably due to the fact that the LASSO penalty cannot achieve consistency and sparsity simultaneously, and the amount of penalty applied here is relatively light in order to maintain adequate estimation accuracy.

It is noteworthy that in all three penalized models, the intercept is negative for component 1 and positive for component 2. Since lower HAQ indicates better health status, patients in the first component are in general healthier than those in the second component. Furthermore, in all three models, the Visit parameter is deleted from component 1, whereas in component 2 it is quite significant. This suggests that the patients' health status does not decline at all for component 1, but declines very severely in component 2. We also observe that, for all three models, the estimate for Baseline HAQ is about 1.5 times larger in component 1 than in component 2. Combined with the parameter estimates for Visit, this implies that the health status of patients in component 1 is primarily determined by their baseline status, and it is less so for component 2. Then, a comparison of the parameter estimates for Skin Score, Shortness of Breath and Raynaud's Phenomenon between the two components also suggest that patients from component 2 suffer greater morbidity than those from component 1. In fact, these three fixed effects are deleted in the Adaptive LASSO and SCAD models. Finally, in all three penalized models, Tender Joint Counts and Digital Ulcers are removed from both mixture components.

Table 7–2: LASSO penalized estimates for the two-component FMLME model.

Variable	Component 1 ($\hat{\pi}_1 = 0.56$, $\hat{\sigma}_1^2 = 0.037$)		Component 2 ($\hat{\pi}_2 = 0.44$, $\hat{\sigma}_2^2 = 0.194$)	
	Fixed (Std. Error)	Random effect SD	Fixed (Std. Error)	Random effect SD
Intercept	-0.064 (0.017)	0.171	0.122 (0.048)	0.423
Visit	–	0.110	0.157 (0.028)	0.217
Male	–	–	-0.093 (0.128)	–
Baseline Age	–	–	–	–
Baseline Duration	–	–	-0.034 (0.046)	–
Baseline HAQ	0.922 (0.016)	–	0.567 (0.041)	–
Skin Score	0.012 (0.014)	0.029	0.078 (0.030)	0.122
Short Breath	0.026 (0.016)	0.085	0.128 (0.031)	0.139
Gastro	–	–	0.033 (0.028)	0.097
Raynaud	0.012 (0.011)	–	0.133 (0.027)	0.120
Tender Joints	–	–	–	–
Ulcers	–	–	–	–

For comparison purposes, we also perform variable selection in the one-component LME model using the method of Bondell et al. (2010) with the Adaptive LASSO penalty, and in

Table 7–3: Adaptive LASSO penalized estimates for the two-component FMLME model.

Variable	Component 1 ($\hat{\pi}_1 = 0.55$, $\hat{\sigma}_1^2 = 0.037$)		Component 2 ($\hat{\pi}_2 = 0.45$, $\hat{\sigma}_2^2 = 0.193$)	
	Fixed (Std. Error)	Random effect SD	Fixed (Std. Error)	Random effect SD
`Intercept`	-0.075 (0.019)	0.192	0.120 (0.046)	0.461
`Visit`	–	0.120	0.158 (0.029)	0.232
`Male`	–	–	-0.058 (0.114)	–
`Baseline Age`	–	–	–	–
`Baseline Duration`	–	–	–	–
`Baseline HAQ`	0.940 (0.017)	–	0.598 (0.038)	–
`Skin Score`	–	–	0.071 (0.025)	0.107
`Short Breath`	–	0.073	0.131 (0.025)	–
`Gastro`	–	–	–	–
`Raynaud`	–	–	0.142 (0.021)	–
`Tender Joints`	–	–	–	–
`Ulcers`	–	–	–	–

Table 7–4: SCAD penalized estimates for the two-component FMLME model.

Variable	Component 1 ($\hat{\pi}_1 = 0.54$, $\hat{\sigma}_1^2 = 0.035$)		Component 2 ($\hat{\pi}_2 = 0.46$, $\hat{\sigma}_2^2 = 0.186$)	
	Fixed (Std. Error)	Random effect SD	Fixed (Std. Error)	Random effect SD
`Intercept`	-0.074 (0.020)	0.196	0.097 (0.052)	0.472
`Visit`	–	0.122	0.143 (0.028)	0.231
`Male`	–	–	–	–
`Baseline Age`	–	–	–	–
`Baseline Duration`	–	–	–	–
`Baseline HAQ`	0.941 (0.017)	–	0.630 (0.041)	–
`Skin Score`	–	–	0.052 (0.027)	0.110
`Short Breath`	–	0.072	0.107 (0.030)	0.127
`Gastro`	–	–	0.013 (0.028)	0.095
`Raynaud`	–	–	0.126 (0.020)	–
`Tender Joints`	–	–	–	–
`Ulcers`	–	–	–	–

the three-component FMLME model using our method with all three penalties. Table 7–5 reports the BIC summary for these models and the two-component FMLME models. From the table, we see that by going from a single LME model to a two-component finite mixture, the BIC is drastically improved. In particular, the penalized two-component FMLME model with Adaptive LASSO penalty improved the BIC by more than 800 units relative to the penalized LME model. Among the three penalized two-component FMLME models, the LASSO penalty yields the model with the largest number of non-zero parameters, which explains why it has the best log-likelihood but the worst BIC. On the other hand, the

Adaptive LASSO FMLME model has the fewest degrees of freedom, and also the best BIC of the two-component FMLME models. Finally, we observe that the gain in log-likelihood is minimal by introducing a third component, and the additional degrees of freedom result in a larger BIC than the corresponding two-component FMLME model. In fact, the best three-component FMLME model still exceeds the two-component Adaptive LASSO model in BIC by 154 units.

On the basis of the above analysis, we conclude that the subjects under study could be divided into two distinct subgroups. In general, patients in the first subgroup enjoy a better health status than those in the second subgroup. And a single LME model would not be able to capture this heterogeneity.

It is worth noting that by including only the participants who had at least four visits, we are more likely to analyze patients that survive longer and thus have better health. If participants who had fewer visits were included, the parameter estimates for visit number and the clinical characteristics would likely increase, but since this data set merely serves to illustrate the proposed methodology, we do not attempt to correct for the survival bias here.

Table 7–5: BIC comparison between the seven penalized models.

Model	Log-likelihood	*df*[*]	Total sample size	BIC
LME(ALASSO)	-1259.41	17	1982	2647.88
Two-component FMLME (LASSO)	-812.31	48	1982	1989.02
Two-component FMLME (ALASSO)	-825.39	24	1982	1832.98
Two-component FMLME (SCAD)	-822.62	33	1982	1895.78
Three-component FMLME (LASSO)	-801.08	93	1982	2308.20
Three-component FMLME (ALASSO)	-811.59	49	1982	1995.18
Three-component FMLME (SCAD)	-799.90	51	1982	1986.98

[*] degree of freedom = number of non-zero parameters

CHAPTER 8
Conclusions and suggestions for future research

In light of the new regularization techniques such as LASSO, SCAD and Adaptive LASSO, we introduced a penalized likelihood approach to simultaneously identify important fixed and random effects in FMLME models, a class of models capable of accounting for both within-subject correlation and between-subject heterogeneity. Theoretical properties of the proposed penalized likelihood estimators were established. The new procedure is shown to consistently select the most parsimonious FMLME model. We devised a computationally efficient nested EM algorithm to perform fixed and random effects estimation and selection. Furthermore, we proposed a data adaptive method to select the regularization parameters and illustrated its use through simulations and a real data example. While traditional best subset methods are often rendered infeasible by their enormous computational cost in many practical situations, our method only requires modest computational resources.

However, several statistical issues related to the new method also deserve further consideration. First, it would be of interest to examine the asymptotic behavior of the tuning parameters chosen by our proposed component-wise BIC criterion. Then, in our methodology, we assumed that each component has a different mean and variance. However, in some applications, there might be prior information indicating that some components may have the same mean but different variances, or different means but the same variance. In such cases, we could modify our maximization strategy accordingly, for example by taking the derivative of (4.4) with respect to the unknown parameters. Another assumption we made was that the number of components K is known a priori, which may not always hold in practice, and leads one to investigate the order estimation problem in FMLME models.

Moreover, our proposed method could be further extended to finite mixtures of generalized linear mixed-effects models, which would permit modeling of repeated count or categorical data with substantial heterogeneity. Finally, it would also be of interest to consider the fixed and random effects selection problem in FMLME models in high-dimensional settings.

References

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csaki (Eds.), *Second international symposium of the information theory.* Budapest: Akademiai Kiado.

Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, *66*, 1069-1077.

Bruce, B., & Fries, J. F. (2003). The stanford health assessment questionnaire: A review of its history, issues, progress, and documentation. *The Journal of Rheumatology*, *30*, 167-78.

Celeux, G., Martin, O., & Lavergne, C. (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, *5*, 1-25.

Chen, Z., & Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, *59*, 762-769.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika*, *82*, 407-410.

Demidenko, E. (2004). *Mixed models: Theory and applications.* New York: Wiley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1-38.

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

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician*, *60*, 19-26.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, *61*, 383-385.

Hennig, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification*, *17*, 273-296.

Ibrahim, J. G., Zhu, H., Garcia, R. I., & Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, *67*, 495-503.

Khalili, A., & Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, *102*, 1025–1038.

Laird, N. M., Lange, N., & Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association*, *82*, 97-105.

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

Lange, N., & Laird, N. M. (1989). The effect of covariance structure on variance estimation in balanced growth-curve models with random parameters. *Bioinformatics*, *84*, 241-247.

Leeb, H., & Pötscher, B. M. (2003). Finite sample distribution of post-model-selection estimators and uniform versus non-uniform approximations. *Econometric Theory*, *19*, 100-142.

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.

Littell, R. C., Pendergast, J., & Natarajan, R. (2000). Modelling covariance structure in the analysis of repeated measures data. *Statistics in Medicine*, *19*, 1793-1819.

Martella, F., Vermunt, J. K., Beekman, M., Westendorp, R. G. J., Slagboom, P. E., & Houwing-Duistermaat, J. J. (2011). A mixture model with random-effects components for classifying sibling pairs. *Statistics in Medicine*, *30*, 3252-3264.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions.* New York: Wiley.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models.* New York: Wiley.

Ng, S. K., McLachlan, G. J., Wang, K., Jones, L. B.-T., & Ng, S.-W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, *22*, 1745-1752.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, *12*, 758-765.

Pu, W., & Niu, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis*, *97*, 733-758.

Scharl, T., Grün, B., & Leisch, F. (2010). Mixtures of regression models for time-course gene expression data: Evaluation of initialization and random effects. *Bioinformatics*, *26*, 370-377.

Schnitzer, M., Hudson, M., Baron, M., & Steele, R. (2011). Disability in systemic sclerosis - a longitudinal observational study. *The Journal of Rheumatology*, *38*, 685-692.

Schwarz, G. E. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461-464.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis* (Second ed.). New York: Wiley.

Snijders, T., & Bosker, R. (2003). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage Publications LTD.

Steele, R. J., & Raftery, A. E. (2010). Performance of Bayesian model selection criteria for Gaussian mixture models. In M.-H. Chen, D. K. Dey, P. Müller, D. Sun, & K. Ye (Eds.), *Frontiers of statistical decision making and bayesian analysis.* New York: Springer.

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

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*, 217-221.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data.* New York: Springer-Verlag.

Wigley, F., & Hummers, L. (2006). Clinical features of systemic sclerosis. In M. Hochberg, A. Silman, J. Smolen, M. Weinblatt, & M. Weisman (Eds.), *Rheumatology.* Amsterdam: Elsevier.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*, 95-103.

Yau, K. K., Lee, A. H., & Ng, A. S. (2003). Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics and Data Analysis*, *41*, 359-366.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418-1429.

KEY TO ABBREVIATIONS

LME model: linear mixed-effects model

FMLME model: finite mixture of linear mixed-effects model

EM algorithm: expectation-maximization algorithm

SSc: systemic sclerosis

HAQ: Health Assessment Questionnaire-Disability Index

AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion

GIC: Generalized Information Criterion

LASSO: Least Absolute Shrinkage and Selection Operator

SCAD: Smoothly Clipped Absolute Deviation

ML: maximum likelihood

REML: restricted maximum likelihood

AR: autoregressive

MSE: mean squared error