University of California
Los Angeles

Reviews of Methods for Variable Selection in Random Effects Model and Some Applications

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Mengxin Tan

2015

Abstract of the Thesis

Reviews of Methods for Variable Selection in Random Effects Model and Some Applications

by

Mengxin Tan

Master of Science in Statistics

University of California, Los Angeles, 2015

Professor Ying Nian Wu, Chair

Linear mixed effects models have been widely used in different disciplines and have become a large research field of Statistics. With the development of science and technology, a large amount of variables are always available to choose for a model and it is necessary to control the numbers of variables to avoid the overfitting problem and use the most efficient way to explain data. Most methods published pay more attention to the selection and estimation of fixed effects but it is meaningful to get a deep insight into variable selection for random effects. Some adjustments have been made in this thesis to obtain the specific methods for variable selection on random effects model based on reviews of some classic or latest methods for variable selection on mixed effects model. These methods and algorithms have been applied on some simulation data and compared through changes on number of subjects and observations. Additionally, these methods have been applied into a real world dataset to study how some effects will influence the democracy index among different countries.

The thesis of Mengxin Tan is approved.

Hongquan Xu

Nicolas Christou

Ying Nian Wu, Committee Chair

University of California, Los Angeles

2015

To my beloved father and mother

without whose help—among so many years—

I couldn't have done this

Table of Contents

List of Tables

CHAPTER 1

Introduction

When analyzing datasets in the fields of Econometrics and Biostatistics, we always need to deal with two particular data types that are quite different from the usual datasets we can use some common statistical methods to study.

One is longitudinal data, also known as panel data or cross-sectional time series data. Longitudinal data is such a dataset that contains multiple entities (countries, states, companies or individuals, etc.) frequently observed at more than two time points or involving in several measurements. For example, when learning what factors may influence a country's GDP, a worldly representative sample of countries may be selected and gathered for their background information over multiple continuous years. This dataset ready for analysis is a classic longitudinal dataset. From datasets of this type, we are able to and need to take advantage of two kinds of information: the differences between different subjects and the differences between different observations at different time points within the same subject.

The other is clustered data. It is a quite simple conception, just a dataset including entities with clustered or grouped labels. For example, when studying income structure, income and other background data are often collected from individuals grouped by different zip codes. There are also datasets naturally grouped. To make a survey on some population information, the observation of each individual are naturally grouped by family or by organization.

Faced with these two types of data, linear mixed effects models proposed by Laird

and Ware (1982) have been widely used in different disciplines and have become a large research field of Statistics with a lot of methods and algorithms proposed and developed on variable estimation and variable selection of linear mixed effects models. Linear mixed effects models are made up of two parts: fixed effects and random effects. Fixed effect includes all the levels of the variable and all the effects of the variable we want to learn have already been in the dataset for analysis. Random effect, in the other way, represents that the variable's levels may be just a random sample selected from a large population. That is to say, for each observation, the form of the distribution for multiple levels of the variable is the same but the parameter of the distribution changes over different observations (Laird and Ware 1982).

What we are interested are quite different in random effects and fixed effects. In fixed effects case, our interest will be in getting the coefficients on the independent variables and what we want to know is the differences between means of the independent variables. While in random effects case, our interest is not really in the specific coefficients' values. What we are really interested in is how the random effects explain the variance of the dependent variable because with this information, it is possible to get it under control. So we want to know more about the variance components of the random effects than their actual values.

In early studies of linear mixed effects model, the numbers of variables used to build models were usually quite small. So the concentration of studies at that time was mostly how to get the solutions to parameters. But with the development of science and technology, it is easier to collect all the background information for a research. That means we are able to obtain nearly all the variables we need related to the model we want to build. To increase the fitness of our model, we tend to include as many useful variables as possible in our model at first. But it is necessary to control the numbers of explanatory variables to avoid the overfitting problem and use the simplest and most efficient way to explain data. So the

next challenge we now face is to include important variables and exclude variables without much influence on the dependent variable. Faced with a large pool of variables, how to deal with variable selection procedure has become a fundamental but quite difficult task in fitting a better model.

A lot of work has been done on variable selection for linear mixed effects model. Lindstrom and Bates (1988) developed a Newton-Raphson algorithm and an EM algorithm for both maximum likelihood and restricted maximum likelihood function of linear mixed effects models. Their work provides a basis of many other algorithms since then to select and estimate parameters in linear mixed effects models. Also the two algorithms had been compared using some real datasets and in most cases, the Newton-Raphson algorithm seems to have the better performance. Based on the EM algorithm and the ECM algorithm, Liu and Rubin (1994) obtained a new algorithm called the ECME algorithm by replacing some CM steps in the ECM algorithm to get a faster convergence. They applied this algorithm in general linear mixed effects models in the next year. Stram and Lee (1994) studied how to test whether variance components of random effects are zero or not in linear mixed effects model by discussing the use of Likelihood Ratio Test. A method proposed by Lin (1997) includes two parts: a global score test to test whether all the variance components of the random effects are zero and individual score tests for each random effect separately. Wang, Song and Zhu (2010) selected and estimated fixed and random effects by adding L1-norm and L2-norm penalties separately for fixed and random effects both in maximum likelihood and restricted maximum likelihood function and solving the maximization problem. Fan and Li (2012) proposed to replace the unknown variance-covariance structure of random effects by proxy matrix in adjusted likelihood functions and select random effects and fixed effects separately. There are also some other methods that select random effects and fixed effects simultaneously.

Most methods available pay more attention to the selection and estimation of

fixed effects. Researchers want to get the most accurate estimates of fixed effects while for random effects, it is just a must to get their variance-covariance matrix estimation so that to better estimate fixed effects. However it is meaningful to get a deep insight into variable selection for random effects. As the number of fixed effects to select is increasing, the size of random effects expands too. Reducing the number of random effects is also necessary to obtain a more efficient and feasible model. Also a random effects model which only contains random effects does exist in the fields of Biostatistics and Econometrics. So the main objective of this thesis is to make brief reviews on variable selection methods already published for linear mixed effects models and do some adjustments on these methods to make them workable for random effects only.

The rest of this thesis will be as follows. Chapter 2 will introduce the basic definitions and theories needed for the variable selection methods reviewed in this thesis. In Chapter 3, I would make reviews on these variable selection methods for random effects only in random effects models and briefly introduce their algorithms. Chapter 4 will carry out simulation studies on these variable selection methods I introduce in Chapter 3 and Chapter 5 will try to perform some of the methods in Chapter 4 on a real world dataset. In the last chapter, some conclusions will be made on these methods used above.

CHAPTER 2

Related Definitions and Theories

2.1 Linear Mixed Effects Model

The form of linear mixed effects models is an extension of a quite common model we use nearly every time we do analysis, linear regression model. The fixed effects in linear mixed effects models are the same both in forms and definitions as the independent variables in linear regression while the random effects are of the same form but their parameters have different definitions from independent variables' coeffecients in linear regression. The standard form of a linear mixed effects model for jth observation in ith subject is

$$y_{ij} = x_{ij}^T \beta + z_{ij}^T b_i + \epsilon_{ij}$$

where y_{ij} is the explained variable from observation j ($j = 1, ..., m_i$ and m_i is the number of observations in subject i) in subject i ($i = 1, ..., n$), x_{ij} is the fixed effects with p covariates $x_{ij1}, x_{ij2}, ..., x_{ijp}$, z_{ij} is the random effects with q covariates $z_{ij1}, z_{ij2}, ..., z_{ijq}$, β is the parameter of fixed effects and is a $p \times 1$ vector, b_i is the parameter of random effects in subject i and is a $q \times 1$ vector and ϵ_{ij} is the error term.

There are some assumptions for the linear mixed effects model which I will use a lot in the following simulation studies:

1.The error term ϵ_{ij} is drawn independently and identically distributed from a normal distribution $N(0, \sigma^2)$ where σ^2 is the error variance.

2.The parameter of random effects b_i is drawn independent and identically dis-

tributed from a multivariate normal distribution $MVN_q(0, \sigma^2 D)$ where σ^2 is the error variance and D is a symmetric matrix.

The matrix form of a linear mixed effects model for each subject i is

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i$$

where $Y_i = (y_{ij})^T, j = 1, ..., m_i$ is a $m_i \times 1$ vector of the explained variable, $X_i = (x_{ij}^T), j = 1, ..., m_i$ is a $m_i \times p$ matrix of the fixed effects, $Z_i = (z_{ij}^T), j = 1, ..., m_i$ is a $m_i \times q$ matrix of the random effects, β is the parameter of fixed effects and is a $p \times 1$ vector, b_i is the parameter of random effects and is a $q \times 1$ vector and $\epsilon_i = (\epsilon_{ij})^T, j = 1, ..., m_i$ is a $m_i \times 1$ vector of the error term.

In this model, $E(Y_i) = X_i\beta$ and $Var(Y_i) = \sigma^2(Z_i D Z_i^T + I_{m_i})$. So the dependent variable Y_i is drawn from a normal distribution $N(X_i\beta, \sigma^2(Z_i D Z_i^T + I_{m_i}))$.

Suppose $V_i = \sigma^2(Z_i D Z_i^T + I_{m_i})$, the maximum log likelihood function for the parameters β, D, σ^2 is

$$l(\beta, D, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|\sigma^2 V_i| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T V_i^{-1}(Y_i - X_i\beta).$$

For $N > p$, Harville (1974) proposed a restricted log likelihood function by modifying the original maximum log likelihood function. The maximum log likelihood function and restricted log likelihood function will be compared in different methods to see whether there exists any big difference between them in selecting and estimating variables. The restricted log likelihood function for the parameters β, D, σ^2 is written as

$$\begin{aligned} l(\beta, D, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|\sigma^{-2} V_i| - \frac{1}{2} log|\sigma^{-2}\sum_{i=1}^{n} X_i^T V_i^{-1} X_i| \\ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - X_i\beta)^T V_i^{-1}(Y_i - X_i\beta) \end{aligned}$$

The linear random effects model is a simplified version of linear mixed effect model and it can be shown as

$$Y_i = \mu + Z_i b_i + \epsilon_i$$

where μ is a $m_i \times 1$ vector of the fixed mean and the values of random effects are standardized so that the means of the random effects can be zero.

In this model, $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2(Z_i D Z_i^T + I_{m_i})$. Still suppose $V_i = \sigma^2(Z_i D Z_i^T + I_{m_i})$, the maximum log likelihood function for the parameters D, σ^2 is

$$l(D, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|\sigma^2 V_i| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i^{-1}(Y_i - \mu)$$

and the restricted log likelihood function for the parameters D, σ^2 is

$$l(D, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|\sigma^{-2} V_i| - \frac{1}{2} log|\sigma^{-2}\sum_{i=1}^{n} J^T V_i^{-1} J| - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i^{-1}(Y_i - \mu)$$

where J is a vector of all ones.

2.2 Newton-Raphson Algorithm

The Newton-Raphson method, also known as Newton Method, is a great technique for finding approximate solutions to the equations' roots numerically.

The approximation we want to obtain is

$$x : f(x) = 0.$$

The key step using iteration is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

We can see that in the iteration step, the derivatives of the equations are required. Sometimes the equations are too complex and the derivatives will be too difficult or even impossible to obtain. In these cases, it will be feasible to replace the derivatives by other approximations.

The Secant Method is a popular way in dealing with this situation as an extension of the Newton-Raphson method when dealing with some difficult derivatives but it

may cause a slower convergence. The slope of a line constructed by two estimates calculated out in two continuous iterations is used to replace the first derivative of the equation.

The key step using iteration here is

$$x_{n+1} = x_n - \frac{f(x_n)}{Q(x_{n-1}, x_n)}$$

where $Q(x_{n-1}, x_n) = (f(x_{n-1}) - f(x_n))/(x_{n-1} - x_n)$.

2.3 EM Algorithm and ECME Algorithm

2.3.1 EM Algorithm

Expectation Maximization (EM) algorithm is an iterative method for parameter estimation especially useful when there exists missing data. It was introduced by Dempster, Laird and Rubin (1977) with proofs by large amounts of applications. EM algorithm makes it possible to obtain maximum likelihood estimates (MLE) and maximum a posteriori estimates (MAP).

The general idea is that there are two steps in each iteration: an expectation (E) step, estimating missing data by estimated parameter values in the maximization step as known parameters and a maximization (M) step, estimating parameters by estimated data in E step as observed data.

Suppose we now have a complete data y and its joint density $f(y; \theta)$. The way we are able to compute the parameter θ is to solve the maximization problem of the maximum likelihood function. The log likelihood function is

$$l(\theta; y) = logL(\theta; y) = logf(y; \theta).$$

The estimate of the parameter θ is the value that maximizes this log likelihood function.

With missing data, we can write the vector y into two parts (y_{obs}, y_{mis}) where y_{obs}

represents the observed part and y_{mis} represents the missing part. Suppose the missing data happens randomly, the joint density function is

$$f(y;\theta) = f(y_{obs}, y_{mis};\theta) = f_1(y_{obs};\theta) f_2(y_{mis}|y_{obs};\theta)$$

where f_1 is the density of the observed data y_{obs} and f_2 is the conditional density of the missing data y_{mis} given y_{obs}. Therefore the log likelihood function of the observed data y_{obs} is

$$l_{obs}(\theta; y_{obs}) = l(\theta; y) - log f_2(y_{mis}|y_{obs};\theta).$$

In usual cases, what we need to obtain the estimate of the parameter θ is to maximize this log likelihood function of the observed data $l_{obs}(\theta; y_{obs})$. We may find it easy to maximize the log likelihood function of the complete data to obtain the estimate of the parameter. But with missing data, the log likelihood function of the complete data can't be estimated and sometimes because of the density structure, solving the maximization problem is too complex. Under this condition, EM algorithm can be used.

E step: Estimate $Q(\theta;\theta^{(t)})$ by

$$Q(\theta;\theta^{(t)}) = E_{\theta^{(t)}}[l(\theta; y)|y_{obs}]$$

M step: Compute $\theta^{(t+1)}$ by

$$Q(\theta^{(t+1)};\theta^{(t)}) \geqslant Q(\theta;\theta^{(t)}).$$

We maximize the log likelihood function of the complete data $l(\theta; y)$ by repeating the E step and the M step iteratively and stop when $\theta^{(t+1)} - \theta^{(t)} < \omega$, where ω is a small quantity.

Lindstrom and Bates (1988) applied the EM algorithm into the linear mixed effect models. The dependent variable Y_i is viewed as the observed data and the parameter of the random effects b_i is viewed as the missing data. The complete data is (Y_i, b_i).

2.3.2 ECME Algorithm

Liu and Rubin (1994) proposed the Expectation Conditional Maximization Either (ECME) algorithm for a faster convergence. The ECME algorithm is an extension of the EM and the ECM algorithm. The difference between EM algorithm and ECM algorithm is that every M step is replaced by Conditional Maximization (CM) step. S steps of conditional maximization are used to make the procedure of maximizing $Q(\theta;\theta^{(t)})$ in M step simpler.

sth CM step: Compute $\theta^{(t+s/S)}$ that

$$Q(\theta^{(t+s/S)};\theta^{t}) \geqslant Q(\theta;\theta^{(t)})$$

where $s = 1, ..., S$ and in next E step iteratively, $\theta^{(t+1)} = \theta^{(t+S/S)}$. The CM step above combined with the same E step as in EM algorithm makes a complete ECM algorithm.

The ECME algorithm adds an "either" by maximizing either $Q(\theta;\theta^{(t)})$ or the log likelihood function of the observed data $l_{obs}(\theta; y_{obs})$ in the original CM step of the ECM algorithm. That is to say, we further divide S steps of conditional maximization into two parts, SQ steps of maximizing $Q(\theta;\theta^{(t)})$ (the same as CM step) and SL steps of maximizing the log likelihood function of the observed data $l_{obs}(\theta; y_{obs})$ where $SQ \cup SL = 1, ..., S$.

2.4 Likelihood Ratio Test and Score Test

2.4.1 Likelihood Ratio Test

Likelihood Ratio Test is a powerful testing method to choose models. It is easy to understand the process of Likelihood Ratio Test. We calculate the likelihood function of each model we assume, compare each model's likelihood function value and choose the model with the biggest likelihood function value.

Suppose L_0 is the likelihood function value of the model from the null hypoth-

esis and L_1 is the likelihood function value of the model from the alternative hypothesis. The likelihood ratio is defined as

$$\lambda = \frac{L_0}{sup(L_1, L_0)}$$

where $0 \leqslant \lambda \leqslant 1$.

In Likelihood Ratio Test, we can use many test statistics, for example Z-Test statistics, F-Test statistics and Chi-Squared Test statistics to test whether $x|\lambda > c$ (accept region) where $0 \leqslant c \leqslant 1$. Though the process of Likelihood Ratio Test seems to be easy, the load of calculation is too heavy to complete by hand so using computer software to perform Likelihood Ratio Test is really necessary.

2.4.2 Score Test

Score Test, also known as the Lagrange Multiplier Test, is another test method which doesn't need any information about the alternative hypothesis.

The process of the Score Test is to first calculate the derivative of the log likelihood function under the null hypothesis (the score) and then calculate the Fisher Information (the score's variance).

The test statistics is computed as follows.

$$S(\theta_0) = \frac{U(\theta_0)^2}{I(\theta_0)}$$

where $U(\theta_0) = \frac{\partial logL(\theta|x)}{\partial \theta}$ and $I(\theta) = -E[\frac{\partial^2}{\partial \theta_2} logL(X;\theta)|\theta]$.

2.5 Shrinkage Penalty

The methods introduced above are all based on the maximum likelihood function. The estimates of parameters from the maximum likelihood function are easy to interpret and have a lot of good properties. But in the case when the number of variables available p is quite large, a smaller set of variables is desired to better

interpret the inner relationship between the explanatory variables X and the explained variable y. To reduce the number of variables or the dimensions of the whole model, it is quite natural to add a penalty on the parameters in the original likelihood function. Thats the motivation and intuition of shrinkage penalty methods.

Instead of maximizing a log likelihood function, we now try to minimize a new function below:

$$M(\theta) = L(\theta|x) + \lambda P(\theta)$$

where $L(\theta|x)$ is a loss function, λ is the tuning parameter and $P(\theta)$ is a penalty function that tends to have a smaller value with less parameters. The penalty function $P(\theta)$ has many different forms, including L_1 penalty, L_2 penalty, the combination of L_1 and L_2 and Smoothed Clipped Absolute Deviance (SCAD) penalty.

We now assume the model

$$y = X\beta + \epsilon$$

where β is a $p \times 1$ parameter vector and $\epsilon \sim (0, \sigma^2)$ to better introduce different penalties below.

2.5.1 L_1 Absolute Penalty (LASSO)

The word LASSO refers to the Least Absolute Shrinkage and Selection Operator. The concept of the LASSO, also known as L_1 penalty, was proposed by Tibshirani (1996). The estimates of parameters are obtained by minimizing

$$(y - X\beta)^T(y - X\beta) \qquad \text{subject to } \sum_{i=1}^{p} |\beta_i| \leqslant t.$$

It can be also written as

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

or also

$$(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1.$$

It tends to get a sparse solution to the parameter β.

2.5.2 L_2 Quadratic Penalty (RIDGE)

The parameter β is further constrained by a ridge penalty, also known as L_2 penalty.

The estimates of parameters are obtained by minimizing

$$(y - X\beta)^T(y - X\beta) \qquad \text{subject to } \sum_{i=1}^{p} \beta_i^2 \leqslant t.$$

It can be also written as

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{i=1}^{p} \beta_i^2$$

or still

$$(y - X\beta)^T(y - X\beta) + \lambda||\beta||_2^2.$$

2.5.3 The Combination of L_1 Penalty and L_2 Penalty (Elastic Net)

Zou and Hastie (2005) developed a new regularized method by a linear combination of the L_1 penalty and L_2 penalty.

The estimates of parameters are obtained by minimizing

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda_1 \sum_{i=1}^{p} \beta_i^2 + \lambda_2 \sum_{i=1}^{p} |\beta_i|$$

or just

$$(y - X\beta)^T(y - X\beta) + \lambda_1||\beta||_2^2 + \lambda_2||\beta||_1.$$

2.5.4 SCAD penalty

To avoid bias problems and obtain continuous solutions, Fan and Li (2001) proposed a new penalty approach. The estimates of parameters are obtained by

minimizing

$$\frac{1}{2}||y - X\beta||^2 + \lambda \sum_{i=1}^{p} p_i(|\beta_i|)$$

or just

$$\frac{1}{2}||y - \hat{y}||^2 + \frac{1}{2}\sum_{i=1}^{p}(z_i - \beta_i)^2 + \lambda \sum_{i=1}^{p} p_i(|\beta_i|)$$

where $z = X^T y$ and $\hat{y} = XX^T y$.

The solution to the parameter is as follows:

$$\hat{\theta} = \begin{cases} sgn(z)(z-\lambda)_+ & \text{when } |z| \leqslant 2\lambda \\ -(n+1)/2 & \text{when } \ 2\lambda \leqslant |z| \leqslant \lambda \\ z & \text{when } |z| > a\lambda. \end{cases}$$

2.5.5 Tuning Parameters

The choice of the tuning parameter λ is quite important. Not only the number of the variables but also the component of regularizations is controlled by the tuning parameter. For example, in L_1 penalty, when the tuning parameter λ approaches 0, the least square estimates are obtained and when the tuning parameter λ approaches $+\infty$, the β^{lasso} is obtained. For each λ, we get one solution path. Cross-validation is a quite popular method in choosing the tuning parameter λ.

CHAPTER 3

Methods and Algorithms

3.1 ECME Algorithm

The following method and algorithm are developed by Schafer (1998).

3.1.1 Model

As shown in Section 2.1, $Y_i \sim N(X_i\beta, \sigma^2(Z_iDZ_i^T + I_{m_i}))$. Suppose now $V_i = (Z_iDZ_i^T + I_{m_i})^{-1}$, $Y_i \sim N(X_i\beta, \sigma^2 V_i^{-1})$.

The likelihood function is given by

$$L_0(\beta, \sigma^2, D) \propto (\sigma^2)^{-\frac{N}{2}} \prod_{i=1}^{n} |V_i|^{\frac{1}{2}} exp\{-\frac{1}{2\sigma^2}(Y_i - X_i\beta)^T V_i (Y_i - X_i\beta)\}$$

where $N = \sum_{i=1}^{n} m_i$.

The restricted maximum likelihood function is given by

$$L_1(\sigma^2, D) \propto (\sigma^2)^{-\frac{N-p}{2}} |\sum_{i=1}^{n} X_i^T V_i X_i| \prod_{i=1}^{n} |V_i|^{\frac{1}{2}} exp\{-\frac{1}{2\sigma^2}(Y_i - X_i\hat{\beta})^T V_i (Y_i - X_i\hat{\beta})\}$$

where $\hat{\beta} = (\sum_{i=1}^{n} X_i^T V_i X_i)^{-1}(X_i^T V_i Y_i)$.

So accordingly the likelihood function of random effects model is given by

$$L_0(\sigma^2, D) \propto (\sigma^2)^{-\frac{N}{2}} \prod_{i=1}^{n} |V_i|^{\frac{1}{2}} exp\{-\frac{1}{2\sigma^2}(Y_i - \mu)^T V_i (Y_i - \mu)\}$$

where $N = \sum_{i=1}^{n} m_i$.

The log likelihood function is shown by

$$l_0(\sigma^2, D) = -\frac{N}{2} log(\sigma^2) + \frac{1}{2} \sum_{i=1}^{n} log(|V_i|) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^T V_i (Y_i - \mu). \quad (3.1)$$

The restricted maximum likelihood function of random effects model is given by

$$L_1(\sigma^2, D) \propto (\sigma^2)^{-\frac{N-p}{2}} |\sum_{i=1}^{n} J_i^T V_i J_i| \prod_{i=1}^{n} |V_i|^{\frac{1}{2}} exp\{-\frac{1}{2\sigma^2}(Y_i - \mu)^T V_i (Y_i - \mu)\}$$

where J_i is a $m_i \times 1$ vector of all ones.

The log restricted likelihood function is shown by

$$\begin{aligned} l_1(\sigma^2, D) = &-\frac{N-p}{2} log(\sigma^2) + log(|\sum_{i=1}^{n} J_i^T V_i J_i|) \\ &+ \frac{1}{2}\sum_{i=1}^{n} log(|V_i|) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i (Y_i - \mu). \end{aligned} \tag{3.2}$$

The parameter of random effects b_i is drawn i.i.d. from a normal distribution whose mean is $E(b_i|Y_i, \sigma^2, D) = \hat{b}_i$ and variance is $Var(b_i|Y_i, \sigma^2, D) = \sigma^2(D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1}$ where

$$\hat{b}_i = (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1} (Y_i - \mu).$$

The conditional mean of the parameter μ is

$$E(\mu|Y_i, \sigma^2, D) = \hat{\mu}$$

where $\hat{\mu} = (\frac{1}{N}\sum_{i=1}^{N} y_{ij})$.

The conditional variance of the parameter μ is

$$Var(\mu|Y_i, \sigma^2, D) = \sigma^2(\sum_{i=1}^{n} J_i^T V_i J_i)^{-1}.$$

So after replacing μ with the estimate $\hat{\mu}$, the distribution of the parameter b_i has a quite complicated mean

$$(D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1} (Y_i - \hat{\mu})$$

and a even more complicated variance

$$\begin{aligned} &\sigma^2[(D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} + (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} \\ &Z_i^T I_{m_i}^{-1} J_i (\sum_{i=1}^{n} J_i^T V_i J_i)^{-1} J_i^T I_{m_i}^{-1} Z_i (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1}]. \end{aligned}$$

3.1.2 Derivatives

This method requires that the inverse of the variance matrix D can be written into

$$D^{-1} = \sum_{i=1}^{k} d_i L_i$$

where d_i is a variance-covariance component in the variance matrix D, L_i is a $q \times q$ symmetric matrix and $k = q(q+1)/2$. L_i is known and is constructed by assigning ones to the locations of the corresponding d_i in the variance matrix D and zeros elsewhere.

3.1.2.1 Maximum Likelihood Function

Suppose $\gamma = \sigma^{-2}$, the equation (3.1) can be simplified into

$$l_0 = \frac{N}{2} log\gamma + \frac{1}{2}\sum_{i=1}^{n} log|V_i| - \frac{\gamma}{2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i (Y_i - \mu).$$

Because $|V_i| = |I_{m_i}|^{-1}|D|^{-1}|(D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1}|$, we now can have the equation (3.1) written into

$$l_0 = \frac{N}{2} log\gamma - \frac{n}{2} log|D| + \frac{1}{2}\sum_{i=1}^{n} log|(D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1}| - \frac{\gamma}{2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i (Y_i - \mu). \tag{3.3}$$

Some important derivatives we prepare for further use:

$$\partial(D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1}/\partial d_j = -(D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} L_j (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1},$$

$$\partial D^{-1}/\partial d_j = L_j,$$

$$\partial V_i/\partial d_j = I_{m_i}^{-1} Z_i (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} L_j (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1}.$$

With these derivatives, the first derivatives of the equations (3.3) can be calculated as

$$\partial l_0/\partial d_j = \frac{1}{2}\sum_{i=1}^{n} tr(D - (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} - \sigma^{-2}\hat{b}_i \hat{b}_i^T) L_j,$$

$$\partial l_0/\partial\gamma = \frac{N}{2}\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)^T V_i(Y_i-\mu)$$

and the second derivatives of the equations (3.3) are

$$\partial^2 l_0/\partial d_j\partial\gamma = -\frac{1}{2}\sum_{i=1}^{n} tr(\hat{b}_i\hat{b}_i^T)L_j,$$

$$\partial^2 l_0/\partial d_j\partial d_k = -\frac{n}{2}trDL_jDL_k + \frac{1}{2}\sum_{i=1}^{n} tr(D^{-1}+Z_i^T I_{m_i}^{-1}Z_i)^{-1}G_j$$
$$(D^{-1}+Z_i^T I_{m_i}^{-1}Z_i)^{-1}G_k + \sigma^{-2}\sum_{i=1}^{n} tr(\hat{b}_i\hat{b}_i^T)L_j(D^{-1}+Z_i^T I_{m_i}^{-1}Z_i)^{-1}L_k,$$

$$\partial^2 l_0/\partial^2\gamma = -N\frac{\sigma^4}{2}.$$

3.1.2.2 Restricted Likelihood Function

The equation (3.2) can be simplified into

$$\begin{aligned} l_1 = \frac{N-p}{2}log\gamma - \frac{n}{2}log|D| + \frac{1}{2}\sum_{i=1}^{n} log|(D^{-1}+Z_i^T I_{m_i}^{-1}Z_i)^{-1}| \\ + \frac{1}{2}log|(\sum_{i=1}^{n} J_i^T V_i J_i)^{-1}| - \frac{\gamma}{2}\sum_{i=1}^{n}(Y_i-\mu)^T V_i(Y_i-\mu). \end{aligned} \tag{3.4}$$

Using the same process as in the section 3.1.2.1, the first derivatives of the equation (3.4) are

$$\partial l_1/\partial d_j = \frac{1}{2}\sum_{i=1}^{n} tr(D-(D^{-1}+Z_i^T I_{m_i}^{-1}Z_i)^{-1} - (D^{-1}+Z_i^T I_{m_i}^{-1}Z_i)^{-1}Z_i^T$$
$$I_{m_i}^{-1}J_i(\sum_{i=1}^{n} J_i^T V_i J_i)^{-1}J_i^T I_{m_i}^{-1}Z_i(D^{-1}+Z_i^T I_{m_i}^{-1}Z_i)^{-1} - \sigma^{-2}\hat{b}_i\hat{b}_i^T)L_j,$$

$$\partial l_1/\partial\gamma = \frac{N-p}{2}\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)^T V_i(Y_i-\mu)$$

and the second derivatives of the equation (3.4) are

$$\partial^2 l_1/\partial^2\gamma = -(N-p)\sigma^4/2,$$

$$\partial^2 l_1/\partial d_j\partial\gamma = -\frac{1}{2}\sum_{i=1}^{n} tr(\hat{b}_i\hat{b}_i^T)L_j/2.$$

From $E(\hat{b}_i) = 0$ and $E(\hat{b}_i\hat{b}_i^T) = \sigma^2(D - (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} - Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1} J_i$ $(\sum_{i=1}^n J_i^T V_i J_i)^{-1} J_i^T I_{m_i}^{-1} Z_i (D^{-1} + Z_i^T I_{m_i}^{-1} Z_i)^{-1})$, we get the expectation of the second derivative of the equation (3.4)

$$E(\partial^2 l_1/\partial d_j \partial d_k) \approx -\frac{1}{2}\sum_{i=1}^{n} tr(D-(D^{-1}+Z_i^T I_{m_i}^{-1} Z_i)^{-1}) L_j (D-(D^{-1}+Z_i^T I_{m_i}^{-1} Z_i)^{-1}) L_k.$$

3.1.3 Iterations

3.1.3.1 Maximum Likelihood Function

σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \mu)^T V_i (Y_i - \mu)$$

and μ can be estimated by

$$\mu = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}.$$

The key iteration steps are shown below:

$$V_i^{(t)} = I_{m_i}^{-1} - I_{m_i}^{-1} Z_i (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1},$$

$$\hat{b}_i^{(t)} = (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1} (Y_i - \mu),$$

$$(\sigma^2)^{(t+1)} = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \mu)^T V_i^{(t)} (Y_i - \mu),$$

$$D^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}((\sigma^{-2})^{(t)} \hat{b}_i^{(t)} \hat{b}_i^{(t)T} + (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1}).$$

3.1.3.2 Restricted Maximum Likelihood Function

σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{1}{N-p}\sum_{i=1}^{n}(Y_i - \mu)^T V_i (Y_i - \mu)$$

and μ can be estimated by

$$\mu = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i} y_{ij}.$$

The key iteration steps are shown below:

$$V_i^{(t)} = I_{m_i}^{-1} - I_{m_i}^{-1} Z_i (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1},$$

$$\hat{b}_i^{(t)} = (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1} (Y_i - \mu),$$

$$A_i^{(t)} = (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} Z_i^T I_{m_i}^{-1} X_i (\sum_{i=1}^{n} J_i^T V_i^{(t)} J_i)^{-1} Z_i^T I_{m_i}^{-1} X_i^T (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1},$$

$$(\sigma^2)^{(t+1)} = \frac{1}{N-p} \sum_{i=1}^{n} (Y_i - \mu)^T V_i^{(t)} (Y_i - \mu),$$

$$D^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} ((\sigma^{-2})^{(t)} \hat{b}_i^{(t)} \hat{b}_i^{(t)T} + (D^{(t)^{-1}} + Z_i^T I_{m_i}^{-1} Z_i)^{-1} + A_i^{(t)}).$$

3.2 Likelihood Ratio Test/Score Test

The following method and algorithm are summarized by Zhang and Lin (2008).

3.2.1 Likelihood Ratio Test

As shown in Section 2.1, $b_i \sim MVN_q(0, \sigma^2 D)$. Suppose D can be interpreted as $D(\eta)$ where η is a vector of variance-covariance components. The objective to select random effects with important and useful information is equivalent to check whose variance components are zero in the variance-covariance matrix $D(\eta)$. When using Likelihood Ratio Test to test every element in the variance-covariance matrix D, some adjustments are needed to make on the distribution which is used to test the statistic because the usual Chi-Squared statistics' distribution can't be used in this case to explain the Likelihood Ratio Test statistic we get.

Self and Liang (1987) proposed an adjusted distribution of the Likelihood Ratio Test statistic. The process is as follows.

First, the Likelihood Ratio Test statistic is still calculated by

$$-2ln\lambda = -2ln\frac{L_0}{sup(L_1, L_0)}$$

where L_0, L_1 are the likelihood functions of different models.

The null hypothesis we use the Likelihood Ratio Test statistic $-2ln\lambda$ to test is

$$H_0 : \eta \in \Omega_0$$

where Ω_0 is the null hypothesis parameter space.

The adjusted distribution Self and Liang proposed is the same as

$$\inf_{C_{\Omega_0}-\eta}(R-x)^T I(\eta)(R-x) - \inf_{C_{\Omega}-\eta}(R-x)^T I(\eta)(R-x)$$

where C_Ω is the cone whose vertex is η_0 and the cone estimates the parameter space Ω, $I(\eta)$ is the Fisher Information on η and R is drawn randomly from a normal distribution $N(0, I^{-1}(\psi_0))$.

The distribution (3.3) can also be written into

$$\inf_{\tilde{C}_0}||\tilde{R}-x||^2 - \inf_{\tilde{C}}||\tilde{R}-x||^2$$

where $\tilde{C} = \{\tilde{x} : \tilde{x} = \Lambda^{1/2}Q^T x \text{ for all } x \in C_\Omega - \eta\}$, $\tilde{C}_0 = \{\tilde{x} : \tilde{x} = \Lambda^{1/2}Q^T x \text{ for all } x \in C_{\Omega_0} - \eta\}$, $\tilde{R}$ is drawn randomly from a normal distribution $N(0, I)$ and $Q\Lambda Q^T$ is a decomposition of the Fisher Information $I(\eta)$.

Stram and Lee (1994) had applied this adjusted Likelihood Ratio Test in linear mixed effects model to test the random effects' variance components.

3.2.2 Score Test

Though the Likelihood Ratio Test introduced above is quite easy to understand and also easy to apply, it is still reasonable to take other test methods into account. The score test is a good choice here by being able to save a lot of work load by deleting the calculation of the likelihood function under alternative hypothesis. Lin (1997) introduced the score test into the field of the linear mixed effects model. The process is as follows.

η, a vector of variance-covariance components, is divided into two parts (η_1, η_2)

where η_1 is of d_1 dimensions and η_2 is of d_2 dimensions. The null hypothesis we use the score test statistic to test is

$$H_0 : \eta_1 = 0.$$

The score is calculated as

$$S_{\eta_1} = n^{-\frac{1}{2}} \frac{\partial l(\eta_1; Y)}{\partial \eta_1}|_{\eta_1=0, \eta_2=\hat{\eta}_2}$$

where $\hat{\eta}_2$ is the maximum likelihood estimate of η_2 under the null hypothesis. The score test statistic is given by

$$T_s = S_{\eta_1}^T H_{\eta_1\eta_1} S_{\eta_1}$$

In equation (3,4), $H_{\eta_1\eta_1}$, the Fisher Information matrix, is given by

$$H_{\eta_1\eta_1} = n^{-1} \tilde{I}_{\eta_1\eta_1}$$

where $I_{\eta_1\eta_1} = I_{\eta\eta} - I_{\eta_1\eta_2}^T I_{\eta_2\eta_2}^{-1} I_{\eta_1\eta_2}$.

3.3 Shrinkage Penalty

The following method and algorithm are developed by Wang, Song and Zhu (2010) and modified by Lin, Pang, and Jiang (2013).

3.3.1 Model

The linear random effects model is a simplified version of linear mixed effect model and it can be shown as

$$Y_i = \mu + Z_i b_i + \epsilon_i$$

where μ is a $m_i \times 1$ vector of the fixed mean and the values of random effects are standardized so that the means of the random effects can be zero.

Suppose $b_i \sim N(0, D)$ and $\epsilon_i \sim N(0, \sigma^2 I_{m_i})$. In this model, $E(Y_i) = \mu$ and $Var(Y_i) = Z_i D Z_i^T + \sigma^2 I_{m_i}$. Now suppose $V_i = Z_i D Z_i^T + \sigma^2 I_{m_i}$, the maximum log likelihood function for the parameters D, σ^2 is

$$l_0(D, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|V_i| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i^{-1}(Y_i - \mu)$$

and the restricted log likelihood function for the parameters D, σ^2 is

$$l_1(D, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|V_i| - \frac{1}{2}log|\sum_{i=1}^{n} J^T V_i^{-1} J| \\ -\frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i^{-1}(Y_i - \mu)$$

where J is a vector of all ones.

3.3.2 Objective Functions

Suppose the parameter of the random effects b_i has a variance matrix D which can be decomposed as LL^T.

To decrease the size of the variables, we add a L2-norm penalty on the likelihood function. The objective function we aim to maximize now is

$$Q(L, \sigma^2) = l(L, \sigma^2) - \lambda \sum_{i=1}^{q} w_i ||L_{(i)}||_2 \tag{3.5}$$

where $w_i = \frac{1}{||L_{(i)}^{(LSE)}||}$ and

$$||L_{(i)}||_2 = \sqrt{L_{i1}^2 + ... + L_{iq}^2}, \qquad i = 1, ..., q.$$

For maximum likelihood function, the estimate of σ^2 can be calculated by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{n}(Y_i - \mu)^T V_i (Y_i - \mu).$$

The $l(L, \sigma^2)$ term in the equation (3.5) we aim to maximize can be written as

$$l_0(D, \sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|V_i| - \frac{1}{2}\sum_{i=1}^{n}(Y_i - \mu)^T V_i^{-1}(Y_i - \mu).$$

For restricted maximum likelihood function, σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{1}{N-p}\sum_{i=1}^{n}(Y_i-\mu)^T V_i(Y_i-\mu).$$

The $l(L,\sigma^2)$ term in the equation (3.5) we aim to maximize can be written as

$$l_1(D,\sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} log|V_i| - \frac{1}{2}log|\sum_{i=1}^{n} J^T V_i^{-1} J| \\ -\frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)^T V_i^{-1}(Y_i-\mu).$$

With the estimate of σ^2, we can update $l(L,\sigma^2)$ into $P(L,\hat{\sigma}^2)$. So the equation (3.5) we aim to maximize is updated into

$$Q(L) = P(L,\hat{\sigma}^2) - \lambda\sum_{i=1}^{q}||L_{(k)}||_2$$

where

$$P(L,\hat{\sigma}^2) = -\frac{1}{2}\sum_{i=1}^{n} log|V_i| - \frac{1}{2}log|\sum_{i=1}^{n} J^T V_i^{-1} J| \\ -\frac{N-p}{2}log(\sum_{i=1}^{n}(Y_i-\mu)^T V_i^{-1}(Y_i-\mu)).$$

3.3.3 Algorithm

To solve this maximization problem, we need to use the Newton-Raphson method and before using this method, we can do some simplifications to this problem.
Suppose $\theta = (\theta_1,...,\theta_k,\theta_{k+1})$ where $k = q(q+1)/2$. $(\theta_1,...,\theta_k)$ is a vector of the variance-covariance components and $\theta_{k+1} = \sigma^2$. Also the vector $(\theta_1,...,\theta_k)$ can be divided into two parts (d,r) where d is a vector of diagonal elements and r is a vector of other elements in the variance-covariance components.
We now can replace the matrix L in the L2 penalty with the vector of diagonal elements d because when the variance d_i equals to zero, the corresponding random effect tends to be not important in this model and we can remove it. The penalty

function now is given by

$$\lambda \sum_{i=1}^{q} w_i \frac{d_i^2}{2|d_i^{(t)}|}.$$

So the equation (3.5) is now calculated as

$$Q(\theta) = P(\theta) - \lambda \sum_{i=1}^{q} w_i \frac{d_i^2}{2|d_i^{(t)}|}.$$

The solution is given by

$$\theta = \underset{\theta}{\operatorname{argmax}}\, Q(\theta)$$

which can be solved by the Newton-Raphson method.

The Newton-Raphson method needs the derivatives of the function. From the function above, the derivatives of the penalty function is quite simple to calculate. The derivatives of the other parts in the restricted likelihood function are given by: Jennrich and Schluchter (1986) providing the calculation process for the likelihood function l_0 and Lindstrom and Bates (1988) providing the calculation process for $-\frac{1}{2} log|\sum_{i=1}^{n} X_i^T V_i^{-1} X_i|$.

- Step 1. Initialize the original values of d as $d^{(0)}$.
- Step 2. Update $d_i^{(t)}$

$$d_i^{(t)} = \underset{d_i}{\operatorname{argmax}}\, P(d) - \lambda \sum_{i=1}^{q} w_i \frac{d_i^2}{2|d_i^{(t)}|}.$$

- Step 3. If $|d_i^t - d_i^{t-1}|$ satisfy that $|d_i^t - d_i^{t-1}| < a$ where a is a quite small quantity such as 10^{-5}, d_i^t is the estimate we want.

  Else, go back to step 2 and continue another iteration.

CHAPTER 4

Simulation Studies

In Chatper 3, there are three methods introduced: the ECME algorithm using both maximum likelihood function and restricted maximum likelihood function, the Likelihood Ratio Test/ score test and the shrinkage penalty method. I apply them in practice with simulation data to see first whether they work well on real data practically and second what properties they may have facing different simulation settings.

I repeat every method in every simulation setting for 100 times and the standard to detect every method's efficiency is to check the percentage of times all random effects are selected correctly as the settings of simulation.

4.1 Simulation Settings

The linear random effects model is

$$Y_i = \mu + Z_i b_i + \epsilon_i$$

where μ is a $m_i \times 1$ vector of the fixed mean and the values of random effects are standardized to have mean 0.

What we need to determine in simulation is as follows.

- Number of subject: n

  which can be viewed as the number of clusters or groups from the original data.

- Number of observations in each subject: m

  which is supposed to be the same for each subject.

- Number of random effects: q

  as the number of fixed effects can be treated as 1 and fixed effects are all one vectors.

- Parameters of random effects: b_i

  which can be sampled from a multivariate normal distribution $MVN(0, \sigma^2 D)$.

  a. D is a $q \times q$ symmetric matrix we can assign values randomly,

  b. σ can be assigned randomly and it is also the variance of the error.

- Values of random effects: Z_i

  which is supposed to be drew from a normal distribution N(0,1) randomly but there is correlation between the observations of the same random effect for the same subject.

4.2 Number of Subjects

Set the number of observations in each subject $m = 10$ and the number of random effects $q = 5$. The model is set as

$$Y_{ij} = \mu + b_{i1} z_{ij1} + b_{i2} z_{ij2} + b_{i3} z_{ij3} + 0 \times z_{ij4} + 0 \times z_{ij5} + \epsilon_{ij}$$

where $i = 1, ..., n$ and $j = 1, ..., 10$. Only the first, second and third random effects are of importance.

Table 3, Table 4, Table 5 and Table 6 shows the first ten results we get from each method. The results from the Likelihood Ratio Test method and the shrinkage penalty method are good enough to justify whether random effects are selected or not. But there is no absolute zero in the values of the variances d_i from the ECME algorithm for us to remove the corresponding variable. It is reasonable

because the ECME algorithm is more likely to be a variable estimation method other than a variable selection method. So we just try to extend the zero standard to 0.01 and 0.005 to see how many variances are below these values.

4.2.1 Small Number of Subjects

Set the number of subject n to be 30.

From Table 1, the number that the ECME method using the likelihood function selects all the random effects of importance is 6 ($d_i < 0.01$) and 1 ($d_i < 0.005$) while the number that the ECME method using the restricted likelihood function selects all the random effects of importance is 8 ($d_i < 0.01$) and 1 ($d_i < 0.005$). The method using the shrinkage penalty select all the random effects of importance for 76 times and the method using the Likelihood Ratio Test select all the random effects of importance for 94 times.

The method using the Likelihood Ratio Test works the best and the percentage it selects the right variables is 94%. The method using the shrinkage penalty also works well with the percentage 76% it selects the right variables. The ECME method has the worst performance but it works a little better using the restricted likelihood function than the likelihood function.

4.2.2 Large Number of Subjects

Set the number of subject n to be 100.

From Table 1, the number that the ECME method using the likelihood function selects all the random effects of importance is 27 ($d_i < 0.01$) and 10 ($d_i < 0.005$) while the number that the ECME method using the restricted likelihood function selects all the random effects of importance is 28 ($d_i < 0.01$) and 10 ($d_i < 0.005$). The method using the shrinkage penalty select all the random effects of importance for 70 times and the method using the Likelihood Ratio Test select all the random

effects of importance for 98 times.

The method using the Likelihood Ratio Test works the best and the percentage it selects the right variables is 98%. The method using the shrinkage penalty also works well with the percentage 70% it selects the right variables. The ECME method has the worst performance but it works a little better using the restricted likelihood function than the likelihood function.

With more subjects, the ECME algorithm has a obviously better performance and the percentage selecting right variables by the Likelihood Ratio Test method increases a little bit by 4%. In opposite, the shrinkage penalty method shows a decreasing trend in performance but it is not obvious.

4.3 Number of Observations in Each Subject

Set the number of subject $n = 30$ and the number of random effects $q = 5$. The model is still set as

$$Y_{ij} = \mu + b_{i1}z_{ij1} + b_{i2}z_{ij2} + b_{i3}z_{ij3} + 0 \times z_{ij4} + 0 \times z_{ij5} + \epsilon_{ij}$$

where $i = 1, ..., n$ and $j = 1, , m$. Only the first, second and third random effects are of importance.

Table 7, Table 8, Table 9 and Table 10 shows the first ten results we get from each method.

4.3.1 Large Number of Observations in Each Subject

Set the number of observations in each subject to be 20.

From Table 2, the number that the ECME method using the likelihood function selects all the random effects of importance is 30 ($d_i < 0.01$) and 11 ($d_i < 0.005$) while the number that the ECME method using the restricted likelihood function selects all the random effects of importance is 35 ($d_i < 0.01$) and 10 ($d_i < 0.005$).

The method using the shrinkage penalty select all the random effects of importance for 78 times and the method using the Likelihood Ratio Test select all the random effects of importance for 96 times.

The method using the Likelihood Ratio Test works the best and the percentage it selects the right variables is 96%. The method using the shrinkage penalty also works well with the percentage 78% it selects the right variables. The ECME method has the worst performance but it works a little better using the restricted likelihood function than the likelihood function.

4.3.2 Small Number of Observations in Each Subject

Set the number of observations in each subject to be 3.

From Table 2, the number that the ECME method using the likelihood function selects all the random effects of importance is 1 ($d_i < 0.01$) and 0 ($d_i < 0.005$) while the number that the ECME method using the restricted likelihood function selects all the random effects of importance is 1 ($d_i < 0.01$) and 0 ($d_i < 0.005$). The method using the shrinkage penalty select all the random effects of importance for 58 times and the method using the Likelihood Ratio Test select all the random effects of importance for 97 times.

The method using the Likelihood Ratio Test works the best and the percentage it selects the right variables is 97%. The method using the shrinkage penalty also works well with the percentage 58% it selects the right variables. The ECME method has the worst performance but it works a little better using the restricted likelihood function than the likelihood function.

With less observations in each subject, the ECME algorithm has a obviously worse performance and the percentage selecting right variables by the Likelihood Ratio Test method decreases sharply by 20%. The Likelihood Ratio Test method doesn't change over the difference of the number of observations.

CHAPTER 5

Real Data Application

5.1 Data Description

The real world dataset I use to apply some of the methods reviewed in Chapter 3 is a subset I draw from the dataset "Oil and Gas Data, 1932-2011" given by Ross (2013). The subset is made up of 72 countries and each country has 10 observations from the year 1966-1996 both randomly drawn from the whole dataset. The standard to clean the dataset is to remove any records with missing data of the explanatory variables.

The response variable we want to study is democracy index scaled increasingly from 1 to 10. We choose six explanatory variables which are the proportion of oil exports in GDP, the proportion of ore and mineral exports in GDP, the proportion of oil exports in GDP lagged 5 years, the proportion of ore and mineral exports in GDP lagged 5 years, the proportion of agriculture outputs in GDP and the proportion of government consumption in GDP. We take the variable country id as the group variable and all the seven variables are treated as random effects. We don't care much about the specific values of the coefficients but how the random effects explain the variance of the democracy levels of countries.

5.2 Analysis Results

Because of the assumption and technique limitation, I just use the shrinkage penalty method and the Likelihood Ratio Test method.

The model we build here is

$$index = \mu + b_1 gov + b_2 oil + b_3 metal + b_4 oil\_5 + b_5 metal\_5 + b_6 agr + \epsilon$$

where each variable's meaning is explained in Table 11.

From Table 12, we can see the estimates of d_i calculated by the shrinkage penalty method. The variances of the variables *medal* and *medal_5* are 7.4×10^{-5} and 1.77×10^{-5}. These values are small enough to approximately take them as zeros. So these two variables can be removed from the whole model. That is to say, we can pick the four variables *oil*, *gov*, *oil_5* and *agr* out of the six variables available. Deep digged into the data we get, the variances of the variables *gov*, *oil_5* and *agr* are quite close to each other in values. So the influences of these three variables are nearly the same on the changes of countries' democracy levels. The variance of the variable *oil* is the largest among all these values. As a result, this variable has the largest influence on the changes of countries' democracy levels.

The p-values from the Likelihood Ratio Test method are also shown in Table 12. The p-value of the variable *metal_5* is 0.802, which is much larger than 0.05. So we can accept the null hypothesis that the variable *metal_5* has no effect on the response variable. So this variable can be removed from the whole model. That is to say, we can pick the five variables *metal*, *oil*, *gov*, *oil_5* and *agr* out of the six variables available.

Comparing these two methods' results, we can find that the variable *metal_5* is removed from the whole model both in these two methods. So the proportion of ore and mineral exports in GDP lagged 5 years doesn't show any importance in this random effects model. But the shrinkage penalty method removes one more variable than the Likelihood Ratio Test method, the variable *metal*. We can see that in this real data case, the shrinkage penalty methods seem to have more power in reducing the size of variables. However it is meaningful to include this variable in the whole model to see its effect.

CHAPTER 6

Conclusion

As most methods published pay more attention to the selection and estimation of fixed effects, there is still space to fill in for random effects only. As a result, I do some adjustments to obtain the specific methods for variable selection on random effects model based on reviews of some classic or latest methods for variable selection on mixed effects model. The methods introduced in Chapter 3 can be summarized as the ECME algorithm, the Likelihood Ratio Test method/ Score Test method and the shrinkage penalty method.

As shown in the simulation studies, the method using the Likelihood Ratio Test works the best and the percentages it selects the right variables are mostly above 95%. The method using the shrinkage penalty also works well with the percentage mostly more than 70% when it selects the right variables. The ECME method has the worst performance but it works a little better using the restricted likelihood function than the likelihood function.

With more subjects, the ECME algorithm has a obviously better performance and the percentage selecting right variables by the Likelihood Ratio Test method increases a little bit while the shrinkage penalty method shows a decreasing trend in performance but it is not obvious. With less observations in each subject, the ECME algorithm has a obviously worse performance and the percentage selecting right variables by the Likelihood Ratio Test method decreases sharply while the Likelihood Ratio Test method doesn't change over the difference of the number of observations.

These methods have been applied into a real world dataset to study how some effects will influence the democracy index among different countries. The shrinkage penalty method and the Likelihood Ratio Test method both work in practice. The shrinkage penalty method seems to have more power in reducing the size of variables.

It is quite easy to develop an algorithm but solving its computing problem will be quite difficult. Several computing methods have been used to make these algorithms introduced in Chapter 3 work. The difficulty of obtaining a new algorithm or method seems to be mainly in how to make it work. In this thesis, there are many simplifications to make the algorithms and methods able to work on real data. For further study, it will be meaningful to take in more computing methods and make these algorithms and methods work in more complex situations.

Tables

Table 1: Summary for Methods, n=30 and n=100

	n=30		n=100	
Method	Number	Per%	Number	Per%
ECME(ML)	6(0.01)/1(0.005)	6%/1%	27(0.01)/10(0.005)	27%/10%
ECME(REML)	8(0.01)/1(0.005)	8%/1%	28(0.01)/10(0.005)	28%/10%
LRT	94	94%	98	98%
SP	76	76%	70	70%

Table 2: Summary for Methods, m=20 and m=3

	m=20		m=3	
Method	Number	Per%	Number	Per%
ECME(ML)	30(0.01)/11(0.005)	30%/11%	1(0.01)/0(0.005)	1%/0%
ECME(REML)	35(0.01)/10(0.005)	35%/10%	1(0.01)/0(0.005)	1%/0%
LRT	96	96%	97	97%
SP	78	78%	58	58%

Table 3: First Ten Results for the ECME Algorithm Using ML and REML, n=30, m=10

the ECME algorithm using ML, n=30, m=10					
Variable	V1	V2	V3	V4	V5
1	11.774121	3.486575	1.0723925	0.006397884	0.018133634
2	8.071108	3.438606	0.8470643	0.062976242	0.011841462
3	10.52151	5.291704	0.5363082	0.030998004	0.052323867
4	6.107029	3.720877	1.4340543	0.021694731	0.044308934
5	9.190416	5.214813	1.0952478	0.026364794	0.040919683
6	10.24079	4.469063	0.9666315	0.021323052	0.018456474
7	7.95517	3.835664	1.0107779	0.058874947	0.01514661
8	9.222144	3.692198	0.8456792	0.013221497	0.031397888
9	8.790593	3.457389	1.0245415	0.030992034	0.027653388
10	7.223828	2.414048	1.5514026	0.018447012	0.001529291
the ECME algorithm using REML, n=30, m=10					
Variable	V1	V2	V3	V4	V5
1	11.773221	3.486386	1.0722625	0.006397194	0.018084041
2	8.070826	3.438674	0.8472209	0.062855425	0.011859939
3	10.521536	5.292018	0.5362319	0.031121815	0.052444055
4	6.106883	3.720888	1.4342319	0.021706441	0.04445823
5	9.190585	5.21491	1.0949823	0.026386388	0.040937326
6	10.24026	4.468902	0.9667764	0.021352548	0.018445056
7	7.954866	3.835685	1.0107492	0.058853673	0.015130113
8	9.221856	3.692581	0.8458299	0.013158043	0.031200363
9	8.790389	3.457647	1.0243527	0.031096662	0.02768564
10	7.223904	2.413866	1.5512954	0.018415584	0.001524064

Table 4: First Ten Results for the ECME Algorithm Using ML and REML, n=100, m=10

the ECME algorithm using ML, n=100, m=10					
Variable	V1	V2	V3	V4	V5
1	7.947413	3.842494	1.0904706	0.02046886	0.021636415
2	8.042949	3.603908	1.0657551	0.002649817	0.00703229
3	9.106388	3.119398	0.9466339	0.038413455	0.01202493
4	9.231878	3.600663	0.7528572	0.002934555	0.010165325
5	9.496526	3.191537	0.8078075	0.005630081	0.010555649
6	7.607902	3.918451	1.0563962	0.005323952	0.01362614
7	9.384627	3.237075	0.9119726	0.009525386	0.010600096
8	10.548298	3.992956	0.8599894	0.018255824	0.002063405
9	7.793814	3.900523	0.7724462	0.04639455	0.012320368
10	8.263566	3.320493	1.241245	0.008388732	0.012352696
the ECME algorithm using REML, n=100, m=10					
Variable	V1	V2	V3	V4	V5
1	7.947398	3.842434	1.0904799	0.020455993	0.021639963
2	8.043	3.603902	1.0657171	0.002648957	0.007031582
3	9.106332	3.119392	0.9466127	0.038448714	0.012040777
4	9.23181	3.600674	0.7528122	0.002935882	0.010157975
5	9.496436	3.191492	0.8078206	0.005641197	0.010581569
6	7.607825	3.918394	1.0564504	0.00531953	0.013639644
7	9.384526	3.23714	0.9119684	0.009519167	0.010598086
8	10.548233	3.992929	0.8599823	0.018253108	0.002068238
9	7.793894	3.900513	0.7724324	0.046412434	0.012330224
10	8.263561	3.320483	1.2412932	0.00840824	0.012354778

Table 5: First Ten Results for the Shrinkage Penalty Method

the shrinkage penalty algorithm, n=30, m=10					
Variable	V1	V2	V3	V4	V5
1	10.15509	3.698694	0.7541604	0	0
2	7.1831776	3.7644246	0.8154478	-0.000834194	0
3	7.212438	2.394458	0.7623177	0	0
4	5.862498	2.699745	0.6664254	0	0
5	13.54439	5.52304	0.8770124	0	0
6	9.19141	4.442611	0.8785027	0	0
7	6.277813	3.115342	0.9953056	0	0
8	5.571921	2.414837	0.9970326	0	0
9	7.34274	3.194443	0.7007853	0	0.01172233
10	7.483506	3.476689	1.208755	0	0

the shrinkage penalty algorithm, n=100, m=10					
Variable	V1	V2	V3	V4	V5
1	8.515331	3.722347	1.046541	0	0
2	8.774047	3.838694	0.8722695	0	0.009801731
3	8.901458	3.938071	0.9319192	0	0
4	6.783247	3.059204	0.9980009	0	0.02124676
5	7.794378	3.664297	0.9086475	0	0
6	6.93084	3.145288	0.9085271	7.93E-05	-4.78E-05
7	8.792047	4.658987	1.158549	0	0
8	8.21777	4.188716	0.8895596	0	0
9	8.795579	3.676892	1.085524	0	0
10	7.525256	4.122904	0.9927831	0	0

Table 6: First Ten Results for the Likelihood Ratio Test Method

the Likelihood Ratio Test method, n=30, m=10					
Variable	V1	V2	V3	V4	V5
1	1.65E-124	4.50E-82	4.07E-15	0.98907439	0.78352609
2	1.64E-50	1.68E-39	9.30E-11	0.96748425	0.99603804
3	4.40E-105	8.20E-75	1.71E-19	0.59599112	0.6907893
4	1.05E-116	1.05E-72	8.38E-21	0.9609925	0.97148564
5	5.94E-91	2.22E-41	1.48E-16	0.46672514	0.40233338
6	7.85E-104	1.88E-53	5.74E-27	0.93784397	0.5887687
7	2.09E-85	5.85E-77	3.44E-27	0.92229142	0.94805514
8	2.12E-98	1.84E-72	4.93E-21	0.32784566	0.84778761
9	2.86E-88	2.71E-61	2.21E-24	0.26039828	0.98444691
10	8.38E-99	3.17E-85	2.27E-38	0.16002618	0.62720037

the Likelihood Ratio Test method, n=100, m=10					
Variable	V1	V2	V3	V4	V5
1	0.00E+00	1.81E-210	9.64E-63	0.39536748	0.741331549
2	0.00E+00	3.00E-265	7.23E-96	0.95976782	0.607013992
3	0.00E+00	1.12E-229	8.60E-97	0.28576693	0.839879124
4	6.21E-292	5.39E-205	6.13E-55	0.32951418	0.393073655
5	0.00E+00	6.35E-194	2.20E-48	0.97605578	0.535955662
6	0.00E+00	5.41E-229	5.81E-67	0.92204478	0.669391705
7	0.00E+00	5.93E-246	4.11E-76	0.53708436	0.990793602
8	0.00E+00	1.06E-255	2.63E-127	0.81000982	0.248343367
9	1.36E-295	2.24E-199	7.91E-69	0.07582091	0.129612348
10	1.61E-270	1.23E-224	1.00E-83	0.97392513	0.714063945

Table 7: First Ten Results for the ECME Algorithm Using ML and REML, n=30, m=20

the ECME algorithm using ML, n=30, m=20					
Variable	V1	V2	V3	V4	V5
1	8.383811	4.251953	1.4023823	0.014274325	0.022726136
2	10.008418	4.267605	1.5473384	0.017002523	0.004666375
3	8.299852	3.435898	0.7596617	0.008791981	0.003927468
4	13.775189	4.488856	0.6963194	0.032576011	0.020482191
5	8.639036	3.321323	0.8233389	0.018810498	0.016582258
6	10.022239	4.051912	1.1103425	0.010014754	0.014261292
7	10.967624	5.860838	1.0072986	0.022381097	0.006927902
8	11.036661	4.918679	1.0325918	0.00611266	0.025074595
9	9.89219	3.972395	1.2424283	0.001503552	0.035329322
10	5.195954	2.977566	0.9749038	0.028877913	0.021108227
the ECME algorithm using REML, n=30, m=20					
Variable	V1	V2	V3	V4	V5
1	8.383752	4.251987	1.4023927	0.014277292	0.022757254
2	10.008463	4.2676	1.5473366	0.017012974	0.004693079
3	8.29974	3.435871	0.7596765	0.008795408	0.003924534
4	13.775149	4.488872	0.6962858	0.032585262	0.020491895
5	8.638959	3.321334	0.8232465	0.018820273	0.016574294
6	10.022353	4.051818	1.1102474	0.010026332	0.01426143
7	10.967686	5.860822	1.0072603	0.022406002	0.006940078
8	11.036624	4.918656	1.0325822	0.006110515	0.025064203
9	9.892148	3.972417	1.242438	0.001501948	0.035368664
10	5.195886	2.977558	0.9749234	0.028872313	0.021124104

Table 8: First Ten Results for the ECME Algorithm Using ML and REML, n=30, m=3

the ECME algorithm using ML, n=30, m=3					
Variable	V1	V2	V3	V4	V5
1	11.494502	5.380629	0.8144257	0.034432345	0.100892444
2	9.301001	3.554087	0.6992064	0.070901507	0.047960296
3	11.894267	5.524333	1.1431493	0.091625876	0.036833305
4	11.370886	4.923334	0.7088352	0.039837167	0.015250724
5	6.682286	3.479442	1.2195698	0.019726175	0.036116221
6	9.342377	4.094098	0.9104179	0.02703775	0.084765008
7	7.546904	3.814029	0.8997437	0.024076578	0.047951729
8	10.675898	3.479413	1.1402276	0.111790776	0.065104865
9	6.507683	6.230968	1.4533118	0.015278548	0.037776529
10	10.677071	3.557221	0.7904649	0.087440036	0.071834676
the ECME algorithm using REML, n=30, m=3					
Variable	V1	V2	V3	V4	V5
1	11.495065	5.380293	0.8140197	0.034390475	0.100895835
2	9.299411	3.553175	0.6990464	0.070646027	0.048161094
3	11.893571	5.521315	1.1410093	0.090500283	0.037098083
4	11.371712	4.922812	0.7083964	0.039927085	0.015342825
5	6.680136	3.480501	1.2183049	0.019529017	0.035978216
6	9.33965	4.094404	0.9097894	0.027057853	0.084478128
7	7.544569	3.813501	0.9007162	0.024047811	0.04753251
8	10.674557	3.47833	1.1386821	0.111991571	0.064917985
9	6.505389	6.231749	1.4533993	0.015420407	0.037671634
10	10.677346	3.557475	0.7899126	0.087796332	0.071777001

Table 9: First Ten Results for the Shrinkage Penalty Method

the shrinkage penalty algorithm, n=30, m=20					
Variable	V1	V2	V3	V4	V5
1	7.626581	3.213082	0.8287373	0.003231832	0
2	8.592947	2.952637	0.6109907	0	0
3	6.283533	2.572192	0.7601286	2.54E-05	0
4	5.793642	2.948914	1.108015	0	0.002683534
5	7.622037	3.430328	0.8057801	0	0
6	7.698335	3.710028	0.9368912	0	0
7	9.720399	3.320429	0.8335052	0	0
8	10.55394	3.495823	0.3944831	0	0.000130373
9	8.183684	4.542988	1.22649	0	0.00156983
10	6.992232	3.059949	0.5512981	0	0
the shrinkage penalty algorithm, n=30, m=3					
Variable	V1	V2	V3	V4	V5
1	4.617253	2.162241	0.6125206	0	0
2	7.95115	2.114707	0.6944613	0	0
3	5.526548	3.950274	0.971626	0.003502358	0.001484028
4	6.282776	3.066445	1.21293	0	0
5	6.066764	2.421973	0.66933	0	0
6	9.85418	6.039	1.257613	0	0
7	4.861325	2.137063	1.019768	1.52E-05	0
8	10.27559	2.841856	0.3745363	0.03421758	0.001832628
9	4.533623	2.462164	0.6462929	0	0
10	8.178193	3.891435	0.6197119	-3.04E-05	0.000290061

Table 10: First Ten Results for the Likelihood Ratio Test Method

the Likelihood Ratio Test method, n=30, m=20					
Variable	V1	V2	V3	V4	V5
1	1.35E-230	1.43E-187	9.49E-69	0.33598241	0.69308171
2	2.66E-266	7.14E-216	3.68E-80	0.26332483	0.3327795
3	1.18E-276	8.99E-198	1.48E-43	0.92291387	0.47004083
4	8.36E-274	7.68E-183	1.72E-67	0.2335256	0.7295692
5	1.24E-240	4.87E-180	1.58E-40	0.32685953	0.38878306
6	1.15E-207	4.11E-151	6.67E-101	0.45251961	0.94547791
7	2.88E-190	2.13E-173	5.20E-65	0.9196948	0.16872392
8	2.39E-219	4.63E-187	6.45E-39	0.95331334	0.36662057
9	2.76E-265	6.68E-170	3.81E-41	0.49567088	0.4994378
10	8.83E-289	9.66E-198	1.31E-68	0.67959479	0.93639657

the Likelihood Ratio Test method, n=30, m=3					
Variable	V1	V2	V3	V4	V5
1	4.01E-37	9.16E-39	1.85E-17	0.787240507	0.838070256
2	3.04E-61	8.36E-36	2.08E-17	0.99984676	0.181376123
3	6.80E-46	4.36E-28	2.74E-14	0.793717227	0.868568609
4	6.33E-62	1.64E-41	1.64E-11	0.564095017	0.190883401
5	6.17E-52	1.32E-36	1.32E-07	0.765409058	0.230740417
6	5.62E-58	3.54E-34	2.04E-02	0.144431076	0.261204686
7	1.12E-55	3.37E-38	2.71E-14	0.777373897	0.116474349
8	1.90E-43	2.68E-33	6.07E-17	0.936952762	0.910134505
9	1.01E-40	1.99E-31	1.45E-13	0.41510491	0.979642077
10	1.46E-44	1.42E-31	2.09E-10	0.782093267	0.71465408

Table 11: Variable Explanation

Variable	Explanation
index	democracy index scaled increasingly from 1 to 10
gov	the proportion of agriculture outputs in GDP
oil	the proportion of oil exports in GDP
metal	the proportion of ore and mineral exports in GDP
oil_5	the proportion of oil exports in GDP lagged 5 years
metal_5	the proportion of oil exports in GDP lagged 5 years
agr	the proportion of agriculture outputs in GDP

Table 12: Summary for Real Dataset

Variable	d_i (Shrinkage)	P-value (LRT)
gov	0.05605723	1.67E-237
oil	0.1337996	0.00E+000
metal	7.40E-005	1.01E-019
oil_5	0.005168929	4.55E-089
metal_5	1.77E-005	8.02E-001
agr	0.00533724	2.56E-193

References

[1] Mary J. Linkdstrom, Douglas M. Bates (1988) "Newton-Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data", *Journal of the American Statistical Association*, 83(404), pp. 1014-1022.

[2] Jennrich RI, Schluchter MD (1986) "Unbalanced repeated-measures models with structured covariance matrices", *Biometrics*, 42(4), pp. 805-20.

[3] Edward Grant (2013) *Selection of Fixed and Random Effects in Linear Mixed Effects Models with Applications to the Trial of Activity in Adolescent Girls*. Unpublished Master Thesis. University of Maryland, College Park.

[4] Bingqing Lin, Zhen Pang and Jiming Jiang (2013) "Fixed and Random Effects Selection by REML and Pathwise Coordinate Optimization", *J Comput Graph Stat.*, 22(2), pp. 341-355.

[5] N. M. Laird and J. H. Ware (1982) "Random Effects Models for Longitudinal Data", *Biometrics*, 38, pp. 963974.

[6] D. A. Harville (1977) "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems", *Journal of the American Statistical Association*, 72, pp. 320-340.

[7] D. A. Harville (1974) "Bayesian Inference for Variance Components Using Only Error Contrasts", *Biometrics*, 61, pp. 383-385.

[8] A. P. Dempster, N. M. Laird and D. B. Rubin (1977) "Maximum Likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society, Series B*, 39, pp. 138.

[9] X. Lin (1997) "Variance Component Testing in Generalized Linear Models with Random Effects", *Biometrika*,84 , pp. 309326.

[10] Sijian Wang, Peter Xuewin Song and Ji Zhu (2010) "Doubly Regularized REML for Estimation and Selection of Fixed and Random Effects in Linear Mixed-Effects Models", *The University of Michigan Department of Biostatistics Working Paper Series*, Working Paper 89, http://biostats.bepress.com/umichbiostat/paper89

[11] J. L. Schafer (1998) "LMM: Some Improved Procedures for Linear Mixed Models", *Software Library for S-PLUS*.

[12] H. Zou and T. Hastie (2005) "Regularization and Variable Selection via the Elastic Net", *Journal of the Royal Statistical Society, Series B (Methodological)*, 67, pp. 301 320.

[13] Y. Fan and R. Li (2012) "Variable Selecion in Linear Mixed Effects Models", *Annals of Statistics*, 40, pp. 20432068.

[14] J. Fan and R. Li (2001) "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties", *Journal of the American Statistical Association*, 96, pp. 134860.

[15] Daowen Zhang, Xihong Lin (2008) "Variance Component Testing in Generalized Linear Mixed Models for Longitudinal/Clustered Data and other Related Topics", *Random Effect and Latent Variable Model Selection*, 192, pp. 19-36.

[16] Michael L. Ross (2013) "Oil and Gas Data, 1932-2011", http://hdl.handle.net/1902.1/20369 UNF:5:dc22RlDasveOTAJvwIjBTA== V2

[17] C. Liu and D. B. Rubin (1994) "The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence", *Biometrika*, 81(4), pp. 633-648.

[18] C. Liu and D. B. Rubin (1995) "ML Estimation of the t Distribution Using EM and Its Extensions, ECM and ECME", *Statistica Sinica*, 5, pp. 19-39.