INFERENCE WITH MISSPECIFIED LINEAR MIXED EFFECTS MODELS

by

YAQIN YANG

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

IN

PROBABILITY AND STATISTICS

CARLETON UNIVERSITY
OTTAWA, ONTARIO

Abstract

This thesis provides an overview of linear mixed effects models commonly used in Biostatistics for analyzing repeated measurements, which include clustered and longitudinal measurements on patients or individuals often considered in clinical studies. To model the correlation structures among repeated measurements, linear mixed models are widely used. Also, these models are used to describe the within-subject errors and between-subject random effects. It is commonly assumed that the vector of random effects in the linear mixed model is normally distributed with a mean vector zero and an unknown variance-covariance matrix. In this thesis, we study the effects of misspecified random effects on the maximum likelihood estimators in linear mixed models. We find that under misspecified random effects distributions the estimators of regression parameters and variance components are generally biased. To reduce the bias in estimation and hence to improve the efficiency of the estimators, skew-normal distributions were suggested by a number of authors. In this thesis, we study the properties of the maximum likelihood estimators under the assumption of skew-normal distributions for the random effects and/or random errors in linear mixed models. We find that when the "true" random effects distributions are skewed, the assumption of a skew-normal distribution for the random effects provides more robust estimators of the model parameters in terms of smaller biases and mean squared errors, as compared to the assumption of a normal distribution for the random effects.

We also study the empirical levels of the likelihood ratio test for testing the significance of the skewness parameters in the skew-normal distribution. Our Monte Carlo

study suggests that the likelihood ratio test provides approximately the correct level of significance under the null hypothesis that the underlying distribution is normal (that is, when the skewness parameter is zero).

As an application of the skew-normal distribution for the random effects and random errors, we present an analysis of some actual longitudinal data on cholesterol levels obtained from the well-known Framingham Heart study.

Acknowledgements

I am truly grateful to my supervisor, Dr. Sanjoy Sinha, for kindly providing guidance throughout the development of this thesis. His directions and comments have been of the greatest help at all times. This work could not have been done without his guidance.

I show my deepest thanks to the authors of my reference paper, Reinaldo B. Arellano-Valle, Heleno Bolfarine and Victor H. Lachos, who made a great contribution of a theoretical basis for skew-normality in linear mixed models.

Moreover, I also like to thank many people in the School of Mathematics and Statistics at Carleton University for their help, continual support, and constructive advice during my graduate education, especially, Shirley Mills and Song Cai.

Finally, I show my sincere appreciation to my father and my mother for supporting me to gain further knowledge in one of the most required fields in the development of several sectors, especially, for their love, encouragement, patience and support. As a return, I intend to dedicate myself to the statistical research and apply them to address real-life issues in my home country.

List of Tables

Contents

Chapter 1

Introduction

In the linear mixed models, the basic assumption about among-subject errors and random effects is normality, or at least symmetric in most cases. But in many situations, the normal assumption cannot estimate the model parameters well in skew-normal scenarios. And in some longitudinal data, it is not frequently the case that the random effects and within subject errors follow a normal distribution.

In this thesis, we consider extending the basic assumption from the normal distribution to the skew-normal distribution for random effects. With this assumption, skew-normal distributions are general situations, and we can also regard normality as a particular case with the skewness parameter $\lambda = 0$. The new assumption provides more possibility to discuss special features of random effects and random errors.

First, we try to study the properties of skew-normal distribution in order to find a proper way to express a skew normal random variable by a combination of normal random variables, if possible, and then we can use the maximum likelihood method to estimate the model parameters. We can convert a skew-normal random variable into a half-normal random variable and a normal random variable. Since it is difficult to derive the joint density function of the vector of outcome variables $\mathbf{Y}$, we first rewrite our model in the form of a hierarchical model and then we derive the marginal distribution of $\mathbf{Y}$. We find the estimators of the regression parameters, variance

components and skewness parameters by maximizing the likelihood function for the given data $\mathbf{Y}$. But there is no explicit solution available for the maximization problem so that the likelihood function has to be maximized numerically. We discuss two special cases of the above general situation, which means that the random effects are assumed to be skew-normal and within-subject errors are assumed to be normal, or the random effects are assumed to be normal and within-subject errors are assumed to be skew-normal. Then we use an EM type algorithm that can give us some convenience to get a maximum value of the log-likelihood.

This thesis is presented as follows. We briefly review the basic concepts of linear mixed models and present definitions of the maximum likelihood estimators in Chapter 2. In Chapter 3, we give the concept of the skew-normal distribution and the construction of the skew-normal distribution, which explains the principle about why we use the combination of density function and distribution function to express the skew-normal distribution. We also discuss some useful properties of the skew-normal distribution, which are really important in the theoretical derivation of the maximum likelihood estimators. In Chapter 4, we describe the use of the EM algorithm which is the theoretical foundation for the ECM algorithm, which is also called the expected conditional maximization. In order to give a comparison, we also introduce the Newton-Raphson method which is one of the most popular computational analysis methods to maximize likelihood functions with normality assumption. In Chapter 5, we run a series of simulations to investigate the average of the estimates, biases, mean squared errors and empirical coverage probabilities. We also study the empirical properties of the likelihood ratio test for assessing the significance of λ at a given level of significance. An illustrative example of the skew-normal analytic method is discussed using longitudinal data on cholesterol levels collected from the famed Framingham Heart Study in Chapter 6. Finally, concluding remarks with some directions for future research are made in Chapter 7.

Chapter 2

Introduction to Linear Mixed Models

The linear mixed model is one of the most popular statistical models, which is widely used in finance, macroeconomics, biology, medicine, and other research fields. Especially in Biostatistics, the linear mixed model is an important technique to explore the relationship between fixed effects, random effects and mixed effects interaction.

There are different types of linear mixed models and different ways of classifying them. One way of classification is according to whether a normality assumption is made or not. As will be seen, normality provides more flexibility in modeling, while models without normality are more robust to violation of distributional assumptions.

Since we focus on the skew-normal assumptions about linear mixed models, we will show more details on the differences between a Gaussian model and a non-Gaussian model. Furthermore, we will illustrate the estimation in linear mixed models based on different methods.

2.1 Linear Models

In a simple linear model, we treat observation vectors as mean vectors and add a difference vector to it. Differences vary from observations to observations, that is to say, we cannot find useful information on these differences. Sometimes, we regard these differences as errors or noises for a model, and these are different from the regression part providing useful information. This technique connects response mean with linear combinations of non-random parameters to make some inferences. For example, in most circumstances, we assume that the data follow a normal distribution, where the general form of a linear model is given by [14]

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.1}$$

In the above equation, $\mathbf{Y}$ is a N-dimensional observed response vector and $\mathbf{X}$ is an observed, fixed and non-random covariate matrix. Also, $\boldsymbol{\beta}$ is unknown and non-random parameter vector that need to be estimated. $\boldsymbol{\epsilon}$ represents random error vector which is usually assumed to be uncorrelated with a zero mean. The main idea of this simple linear model is that the mean vector of $\mathbf{Y}$ can be represented as a linear combination of the covariate vectors.

2.2 Random Effects Model

As for random effects models, they usually connect observations vector with intercept vector and some levels of random groups from the population. Here, we should note the difference between fixed effects and random effects, although they can have the same expression in some cases. The random effects are randomly chosen levels from all population levels. But the fixed effects are different levels of treatment effects that we are interested in. And they are decided before we conduct the experiments. In order to illustrate it clearly, suppose we want to investigate health conditions of

retired people at 15 different communities in the whole city. The model could be [14]

$$E[y_{ij}] = \mu + \beta_i. \tag{2.2}$$

Here we denote the observation from the jth retired person of ith community by y_{ij}, i=1,2,...,15, which represent 15 different communities chosen from all communities in the whole city. Here we should note that β_i is not a fixed effect but a random effect, which means that the ith community is not pre-decided in conducting the experiment. The community i is the one community which is randomly chosen and numbered i in the health investigation. And since these randomly chosen communities aim to be regarded as a substitute of the population communities in the whole city, we can use this sample information to estimate the population. This is a common property of random effects, which acts as the foundation for deducing about the population. [14] Thus β_i is a random variable, of which the data can give useful information to deduce the variance of other random variables. By the way, μ is a general mean.

In practice, random effects have more different properties than fixed effects since we assume that [14]

$$\beta_i \sim i.i.d.(0, \sigma_\beta^2), \tag{2.3}$$

which means that [14]

$$E[\beta_i] = 0,$$

$$var[\beta_i] = E[(\beta_i - E[\beta_i])^2] = E[\beta_i^2] = \sigma_\beta^2.$$

Here σ_β^2 is the within-groups covariance, which measures relativity of every two observations in one group.

In fact, random effects models are used less than linear mixed models in many actual problems, since linear mixed models not only take random effects into account but also take the mixture of fixed and random factors into account comprehensively.

2.3 Linear Mixed Models

The general form of linear fixed models is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}. \tag{2.4}$$

Here, $\mathbf{X}$ is the N-dimensional design matrix of fixed effects, and $\boldsymbol{\beta}$ is a coefficients vector of fixed effects. At the same time, $\mathbf{Z}$ represents the design matrix for random effects, and $\mathbf{b}$ is coefficients vector of random effects in matrix form. $\boldsymbol{\epsilon}$ is the vector of random errors.

In order to illustrate the property of the variance-covariance matrix, we give a simple form of linear mixed models as an example here. Assume that the response y_{ij} is described as a function of the covariate x_{ij} by the linear mixed model

$$y_{ij} = \beta_0 + x_{ij}\beta_1 + b_i + \epsilon_{ij}, \quad (i = 1, \ldots, m, j = 1, \ldots, n_i, N = \textstyle\sum_{i=1}^{m} n_i) \tag{2.5}$$

where random effects b_i's are independent normal variables $N(0, \sigma_b^2)$, random errors ϵ_{ij}'s are independent normal variables $N(0, \sigma_\epsilon^2)$, and b_i and ϵ_{ij} are independent of each other.

For this model, the marginal variance is obtained as

$$\begin{aligned} var(y_{ij}) &= var(\beta_0 + x_{ij}\beta_1 + b_i + \epsilon_{ij}) \\ &= var(\beta_0 + x_{ij}\beta_1) + var(b_i + \epsilon_{ij}) \\ &= 0 + var(b_i + \epsilon_{ij}) \\ &= var(b_i) + var(\epsilon_{ij}) \\ &= \sigma_b^2 + \sigma_\epsilon^2. \end{aligned}$$

For the covariance of the inter-group, we have

$$\begin{aligned} cov(y_{ij}, y_{ij'}) &= cov(\beta_0 + x_{ij}\beta_1 + b_i + \epsilon_{ij}, \beta_0 + x_{ij'}\beta_1 + b_i + \epsilon_{ij'}) \\ &= cov(\beta_0 + x_{ij}\beta_1, \beta_0 + x_{ij'}\beta_1) + cov(b_i + \epsilon_{ij}, b_i + \epsilon_{ij'}) \\ &= 0 + cov(b_i + \epsilon_{ij}, b_i + \epsilon_{ij'}) \\ &= var(b_i) + cov(b_i, \epsilon_{ij'}) + cov(\epsilon_{ij}, b_i) + cov(\epsilon_{ij}, \epsilon_{ij'}) \\ &= \sigma_b^2. \end{aligned}$$

For the covariance of the intra-group, we have

$$\begin{aligned} cov(y_{ij}, y_{i'j}) &= cov(\beta_0 + x_{ij}\beta_1 + b_i + \epsilon_{ij}, \beta_0 + x_{i'j}\beta_1 + b_{i'} + \epsilon_{i'j}) \\ &= cov(\beta_0 + x_{ij}\beta_1, \beta_0 + x_{i'j}\beta_1) + cov(b_i + \epsilon_{ij}, b_{i'} + \epsilon_{i'j}) \\ &= 0 + cov(b_i + \epsilon_{ij}, b_{i'} + \epsilon_{i'j}) \\ &= cov(b_i, b_{i'}) + cov(b_i, \epsilon_{i'j}) + cov(\epsilon_{ij}, b_{i'}) + cov(\epsilon_{ij}, \epsilon_{i'j}) \\ &= 0. \end{aligned}$$

Similarly, $cov(y_{ij}, y_{i'j'}) = 0$.

We can also present model (2.7) in the matrix form (2.6), with $\mathbf{Y} = (\mathbf{Y}_1', \ldots, \mathbf{Y}_i', \ldots, \mathbf{Y}_m')'$, where the ith response vector $\mathbf{Y}_i$ with n_i dimensions is modeled as a function of the design matrix $\mathbf{X}_i$ by [14]

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \quad (i = 1, \ldots, m) \tag{2.6}$$

The marginal mean response is $E(\mathbf{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ and the marginal covariance matrix is

$$\mathbf{V}(\mathbf{Y}_i) = \mathbf{V}_i = \begin{pmatrix} \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \cdots & \sigma_b^2 \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma_\epsilon^2 \end{pmatrix}.$$

The marginal variance-covariance matrix for the whole response vector $\mathbf{Y}$ is given by

$$\mathbf{Var}(\mathbf{Y}) = \begin{pmatrix} \mathbf{V}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{V}_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathbf{V}_m \end{pmatrix} = \mathbf{V}.$$

In fixed models and linear mixed models, we are generally interested in estimating the fixed effects regression parameters or some estimable linear functions of these parameters, particularly to find differences between the levels of any given treatments. When random effects become the part of a model, we often want to estimate the variance of the within-subject random errors and between-subject random effects, and test the significance of the fixed effects parameters as well as the variances of the random effects.

In practice, we have different kinds of linear models, and different ways to classify these linear models. Here we want to choose a way to classify linear models according to whether they are under a normality assumption or not. We can see it in later chapters that the assumed normal distribution could give us more flexibility when we fit the model, while models with assumed skew-normal distribution are more robust to violation of distributional assumptions. [18]

Under linear mixed models with the assumed normal distribution, parameters of the covariance-variance matrix can be estimated by a suitable method, such as the maximum likelihood method. We describe below the use of a normal mixed model for analyzing longitudinal data.

Longitudinal model: Since these types of models are often used in the analysis of longitudinal data, we often call them longitudinal models, which are also a class of linear mixed models. [12] A characteristic of longitudinal models is that we usually divide the observation data into several groups with a random effect (or a vector

of random effects). In practice, we combine these groups of different individuals appeared in the longitudinal study. What's more, there may be serial correlations within each group as we have derived in the matrix form. Another characteristic of the longitudinal models is that there are often time-dependent covariates, which may appear either in $\mathbf{X}$ or $\mathbf{Z}$, just as considered in the simulation study in Chapter 5.

Following a introduced measurement of uncertainty of estimated function [10], a general longitudinal model may be expressed as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i, \tag{2.7}$$

where $\mathbf{Y}_i$ represents the n_i-dimensional vector of observations from the ith individual; $\mathbf{X}_i$ and $\mathbf{Z}_i$ are known design matrices; $\boldsymbol{\beta}$ is an unknown vector of regression coefficients; $\mathbf{b}_i$ is a vector of random effects; and $\boldsymbol{\epsilon}_i$ is a vector of random errors. It is assumed that $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are independent with $\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{D}_i)$, $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \mathbf{R}_i)$, where the covariance matrices $\mathbf{D}_i$ and $\mathbf{R}_i$ are known up to a vector $\boldsymbol{\theta}$ of variance components. [10] Here the assumptions of covariance-variance matrices of random effects and random errors are in general form, and we should note that $\mathbf{D}_i$ and $\mathbf{R}_i$ vary according to different groups. To be specific, we can have the following general variance-covariance matrix for the n_i-dimensional response vector $\mathbf{Y}_i$: [10]

$$\mathbf{Var}(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{D}_i\mathbf{Z}_i^T + \mathbf{R}_i = \mathbf{V}_i$$

We also should note that the variance-covariance matrix of the random effects, $\mathbf{D}_i$, is not restricted to a certain scalar times the identity matrix, in fact, it can be any symmetric matrix.

The longitudinal model with the normality assumption may also be extended to the skew-normality assumption. The typical linear mixed model with skew-normality assumption is such that $\mathbf{Y}_1, ..., \mathbf{Y}_m$ are independent and have the same form $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i$. Furthermore, for each i, $\mathbf{b}_i$ and $\boldsymbol{\epsilon}_i$ are uncorrelated with $E(\mathbf{b}_i) = \mathbf{0}$,

$Var(\mathbf{b}_i) = \mathbf{G}_i$; $E(\boldsymbol{\epsilon}_i) = \mathbf{0}$, $Var(\boldsymbol{\epsilon}_i) = \mathbf{R}_i$. [10] Alternatively, the independence of $\mathbf{Y}_i$, $1 \leq i \leq m$ may be replaced by that of $(\mathbf{b}_i, \boldsymbol{\epsilon}_i)$, $1 \leq i \leq m$. Under linear mixed models without the assumption of normality, the random effects and errors are assumed to be independent, or simply uncorrelated, but their distributions are not assumed to be normal. All the other assumptions are the same as in the normality assumption case. Again, in this case, the distribution of $\mathbf{Y}$ may not be fully specified up to a set of parameters, or, even if it can be fully specified up to a set of parameters, it may not have a closed-form expression. [10] We observe that in Chapter 5 for the simulation study, and adopt a different method to analyze.

Marginal model Alternatively, a linear mixed model with the assumption of normality may be expressed by its marginal distribution. To see this, note that under the linear mixed model and normality, we have

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \tag{2.8}$$

where $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$. As for the longitudinal model, one may assume that $\mathbf{Y}_1, \cdots, \mathbf{Y}_m$ are independent with $\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$, where $\mathbf{V}_i = \mathbf{R}_i + \mathbf{Z}_i\mathbf{D}_i\mathbf{Z}_i^T$. It is clear that the model can also be expressed with $\mathbf{R} = diag(\mathbf{R}_1, ..., \mathbf{R}_m)$ and $\mathbf{G} = diag(\mathbf{D}_1, ..., \mathbf{D}_m)$, whose variance-covariance matrix is given by [14]

$$Var(\mathbf{Y}) = \begin{pmatrix} \mathbf{Z}_1\mathbf{D}_1\mathbf{Z}_1^T + \mathbf{R}_1 & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2^T + \mathbf{R}_2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathbf{Z}_{m-1}\mathbf{D}_{m-1}\mathbf{Z}_{m-1}^T + \mathbf{R}_{m-1} & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{Z}_m\mathbf{D}_m\mathbf{Z}_m^T + \mathbf{R}_m \end{pmatrix}$$

and $\mathbf{X}$ and $\mathbf{Z}$ are defined as before.

A disadvantage of the marginal model is that the random effects are not explicitly

defined. In many cases, these random effects have practical meanings, and the inference about them may be of interest. [15] But the marginal model perhaps is the most general model among all types. Under a marginal model, it is assumed that $\mathbf{Y}$, the vector of observations, satisfies $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $Var(\mathbf{Y}) = \mathbf{V}$, where $\mathbf{V}$ is specified up to a vector $\boldsymbol{\theta}$ of variance components. The random effects are not present in the marginal model. Therefore, the model has the disadvantage of not being suitable for inference about the random effects. But if the estimate of a random effect is our main interest, we may try another suitable method, just as we will show in the simulation study. Also, because the model is so general that not many assumptions are made, it is often difficult to assess (asymptotic) properties of the estimators.

By not fully specifying the distribution, a linear mixed model without the assumption of normality may be more robust to violation of distributional assumptions. On the other hand, methods of inference that require specification of the parametric form of the distribution, such as maximum likelihood, may not apply to such a case. The inference about both assumptions of normality and skew-normality of linear mixed models is discussed in the rest of this thesis.

Hierarchical models From a Bayesian point of view, we can separate a linear mixed model into three-stage hierarchy. At the beginning, the distribution of the observations is defined when the random effects is given. In the second stage, the distribution of the random effects given the model parameters is defined. Finally, a prior distribution is assumed for the parameters. These stages may be specified as follows under normality. Let $\boldsymbol{\theta}$ represent the vector of variance components involved in the model. Then, we have [14]

$$\mathbf{y} \mid \mathbf{b}, \boldsymbol{\theta}, \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R}),$$

$$\mathbf{b} \mid \boldsymbol{\theta} \sim N(\mathbf{0}, \mathbf{D}),$$

and

$$(\boldsymbol{\beta}, \boldsymbol{\theta}) \sim \omega(\boldsymbol{\beta}, \boldsymbol{\theta})$$

where $\mathbf{R} = \mathbf{R}(\boldsymbol{\theta})$, $\mathbf{D} = \mathbf{D}(\boldsymbol{\theta})$, and ω is a known distribution. In many cases, it is assumed that $\omega(\boldsymbol{\beta}, \boldsymbol{\theta}) = \omega_1(\boldsymbol{\beta})\omega_2(\boldsymbol{\theta})$, where $\omega_1 = N(\boldsymbol{\beta}_0, \mathbf{D})$ with both $\boldsymbol{\beta}_0$ and $\mathbf{D}$ known, and ω_2 is a known distribution. [14]

2.4 Maximum likelihood estimation

Standard methods of estimation in linear mixed models with assumptions of normality or skew-normality are the maximum likelihood (ML) or other methods. In this section, we discuss the maximum likelihood method. Although the estimation of the variance components in a linear mixed model was not easy to handle computationally in the old days, the estimation of the fixed effects given the variance components is straightforward. [16]

Point estimation Under a linear mixed model with the assumption of normality, the distribution of $\mathbf{Y}$ is given by (2.9), which has a joint pdf

$$f(\mathbf{Y}) = \frac{1}{(2\pi)^{N/2} \mid \mathbf{V} \mid^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\} \tag{2.9}$$

where $N = \sum_{i=1}^{m} n_i$ is the dimension of $\mathbf{Y}$. Thus, the log-likelihood function is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = c - \frac{1}{2} log(\mid \mathbf{V} \mid) - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{2.10}$$

where $\boldsymbol{\theta}$ represents the vector of all the variance components (involved in $\mathbf{V}$), and c is a constant.

Here we will introduce some useful results of matrix differentiation. If $\mathbf{A}$ is a matrix whose elements are functions of $\boldsymbol{\theta}$, a real-valued variable, then $\partial\mathbf{A}/\partial\boldsymbol{\theta}$ represents the matrix whose elements are the derivatives of the corresponding elements of $\mathbf{A}$

with respect to $\boldsymbol{\theta}$. For example, if [14]

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \tag{2.11}$$

then

$$\frac{\partial \mathbf{A}}{\partial \theta} = \begin{pmatrix} \partial a_{11}/\partial\theta & \partial a_{12}/\partial\theta \\ \partial a_{21}/\partial\theta & \partial a_{22}/\partial\theta \end{pmatrix}. \tag{2.12}$$

If $\mathbf{a} = (a_i)_{1\leq i\leq k}$ is a vector whose components are functions of $\boldsymbol{\theta} = (\boldsymbol{\theta}_i)_{1\leq j\leq l}$, a vector-valued variable, then $\partial\mathbf{a}/\partial\boldsymbol{\theta}'$ is defined as the matrix $(\partial\mathbf{a}_i/\partial\boldsymbol{\theta}_i)_{1\leq i\leq k,1\leq j\leq l}$. Similarly, $\partial\mathbf{a}/\partial\boldsymbol{\theta}'$is defined as the matrix $(\partial\mathbf{a}/\partial\boldsymbol{\theta}')'$. The following are some useful results. [14]

(i) (Inner-product) If $\mathbf{a}$, $\mathbf{b}$, and $\boldsymbol{\theta}$ are vectors, then

$$\frac{\partial(\mathbf{a}^T\mathbf{b})}{\partial\boldsymbol{\theta}} = \left(\frac{\partial\mathbf{a}^T}{\partial\boldsymbol{\theta}}\right)\mathbf{b} + \left(\frac{\partial\mathbf{b}^T}{\partial\boldsymbol{\theta}}\right)\mathbf{a}. \tag{2.13}$$

(ii) (Quadratic form) If $\mathbf{x}$ is a vector and $\mathbf{A}$ is a symmetric matrix, then

$$\frac{\partial}{\partial\mathbf{x}}\mathbf{x}^T\mathbf{A}\mathbf{x} = 2\mathbf{A}\mathbf{x}. \tag{2.14}$$

(iii) (Inverse) If the matrix $\mathbf{A}$ depends on a vector $\boldsymbol{\theta}$ and is nonsingular, then, for any component $\boldsymbol{\theta}_i$ of $\boldsymbol{\theta}$,

$$\frac{\mathbf{A}^{-1}}{\partial\boldsymbol{\theta}_i} = -\mathbf{A}^{-1}\left(\frac{\partial\mathbf{A}}{\partial\boldsymbol{\theta}_i}\right)\mathbf{A}^{-1}. \tag{2.15}$$

(iv) (Log-determinant) If the matrix $\mathbf{A}$ above is also positive definite, then, for any component $\boldsymbol{\theta}_i$ of $\boldsymbol{\theta}$,

$$\frac{\partial}{\partial\boldsymbol{\theta}_i} log(\mid \mathbf{A} \mid) = tr\left(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\boldsymbol{\theta}_i}\right). \tag{2.16}$$

By differentiating the log-likelihood with respect to the parameters, we obtain the

following score functions for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$:

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} - \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \tag{2.17}$$

$$\frac{\partial l}{\partial \boldsymbol{\theta}_r} = \frac{1}{2}\left\{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_r}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) - tr\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_r}\right)\right\} \tag{2.18}$$

where $r = 1, \cdots, q$, $\boldsymbol{\theta}_r$ is the rth component of $\boldsymbol{\theta}$, which has the dimension q. The standard procedure of finding the ML estimators (MLE), is to solve the ML equations $\partial l/\partial \boldsymbol{\beta} = 0$, $\partial l/\partial \boldsymbol{\theta} = 0$. However, finding the solutions may not be the end of the story. In other words, the solutions to (2.18) and (2.19) may or may not be the MLE. Let p be the dimension of $\boldsymbol{\beta}$. For simplicity, we assume that $\mathbf{X}$ is of full (column) rank; that is, $rank(\mathbf{X}) = p$. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ be the MLE. From (2.18) one obtains [14]

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{2.19}$$

where $\hat{\mathbf{V}} = \mathbf{V}(\hat{\boldsymbol{\theta}})$, that is, $\mathbf{V}$ with the variance components involved replaced by their MLE. Thus, once the MLE of $\boldsymbol{\theta}$ is found, the MLE of $\boldsymbol{\beta}$ can be calculated by the closed-form expression (2.20). As for the MLE of $\boldsymbol{\theta}$, by (2.18) and (2.19) it is easy to show that it satisfies [14]

$$\mathbf{y}^T\mathbf{P}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_r}\mathbf{P}\mathbf{y} = tr\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_r}\right), \tag{2.20}$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}. \tag{2.21}$$

Thus, one procedure is to first solve (2.21) for $\hat{\boldsymbol{\theta}}$ and then compute $\hat{\boldsymbol{\beta}}$ by (2.20). In the special case of linear mixed models with the original form of variance components,

we have $Var(\mathbf{Y}_i) = \mathbf{V}_i = \sigma_e^2 \mathbf{I}_{n_i} + \sigma_b^2 \mathbf{Z}_i \mathbf{Z}_i^T$, whose matrix form is

$$\mathbf{V}_i = \begin{pmatrix} \sigma_b^2 + \sigma_\epsilon^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_\epsilon^2 & \cdots & \sigma_b^2 \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_b^2 & \sigma_b^2 & \cdots & \sigma_b^2 + \sigma_\epsilon^2 \end{pmatrix}_{n_i \times n_i}$$

For the whole response vector $\mathbf{Y}$, we have

$$\mathbf{Var}(\mathbf{Y}) = \begin{pmatrix} \mathbf{V}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{V}_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathbf{V}_m \end{pmatrix}$$

Thus, $\partial \mathbf{V}_i / \partial \sigma_e^2 = \mathbf{I}_n$, $\partial \mathbf{V}_b / \partial \sigma_b^2 = \mathbf{Z}_i \mathbf{Z}_i^T$, $1 \le i \le m$. That is,

$$\partial \mathbf{V}_i / \partial \sigma_e^2 = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}_{n_i \times n_i},$$

$$\partial \mathbf{V}_i / \partial \sigma_b^2 = \begin{pmatrix} z^2 & z^2 & \cdots & z^2 & z^2 \\ z^2 & z^2 & \cdots & z^2 & z^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z^2 & z^2 & \cdots & z^2 & z^2 \\ z^2 & z^2 & \cdots & z^2 & z^2 \end{pmatrix}_{n_i \times n_i}$$

Thus we can have that

$$\partial \mathbf{V}/\partial \sigma_e^2 = \begin{pmatrix} \mathbf{I}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{n_2} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_{n_{m-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{I}_{n_m} \end{pmatrix},$$

and

$$\partial \mathbf{V}/\partial \sigma_b^2 = \begin{pmatrix} \partial \mathbf{V}_1/\partial \sigma_b^2 & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \partial \mathbf{V}_2/\partial \sigma_b^2 & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \partial \mathbf{V}_{m-1}/\partial \sigma_b^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \partial \mathbf{V}_m/\partial \sigma_b^2 \end{pmatrix}.$$

Asymptotic covariance matrix: Under suitable conditions, the MLE is consistent and asymptotically normal with the asymptotic covariance matrix equal to the inverse of the Fisher information matrix [25]. Let $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$. Then, under regularity conditions, the Fisher information matrix has the following expressions, [14]

$$Var\left(\frac{\partial l}{\partial \boldsymbol{\psi}}\right) = -E\left(\frac{\partial^2 l}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T}\right) \tag{2.22}$$

By (2.18) and (2.19), further expressions can be obtained for the elements of (2.23). For example, assuming that $\mathbf{V}$ is twice continuously differentiable (with respect to the components of $\boldsymbol{\theta}$), it can be shown that [14]

$$E\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right) = -\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}, \tag{2.23}$$

$$E\left(\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}_r}\right) = 0, \tag{2.24}$$

$$E\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s}\right) = -\frac{1}{2} tr\left(V^{-1}\frac{\partial V}{\partial \boldsymbol{\theta}_r}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \boldsymbol{\theta}_s}\right). \tag{2.25}$$

It follows that (2.23) does not depend on $\boldsymbol{\beta}$, and therefore may be denoted by $I(\boldsymbol{\theta})$, where

$$I(\boldsymbol{\theta}) = -\mathbf{E}\begin{pmatrix} \frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} & \frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\theta}^T} \\ \frac{\partial^2 l}{\partial \boldsymbol{\beta}^T\partial \boldsymbol{\theta}} & \frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T} \end{pmatrix} = \begin{pmatrix} \frac{\partial \boldsymbol{\mu}^T}{\partial \boldsymbol{\beta}}\mathbf{V}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\left\{tr(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_r}\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \theta_s})\right\} \end{pmatrix}$$

In Chapter 4, we will illustrate the EM algorithm and Newton-Raphson method to get the maximum likelihood estimates computationally, where the above score functions and information matrices play an important role.

Chapter 3

Skew-normal Distribution

3.1 Overview of Skew-normal Distributions

In this section, we will talk about definitions and properties of the skew-normal distribution, and give the marginal distribution of the general case which involves two skew-normal assumptions. Firstly, we will illustrate the construction of skew-normal distribution, which explains the reason why we express skew normal distribution as the combination of density function and distribution function. Next, we will introduce some useful properties of skew-normal distribution, which are really important in the following theoretical derivation. And we also derive the marginal distribution of $\mathbf{Y}$ of the general situation. Finally, we will give two special cases of the above general situation, that is, we assume that the distribution of the random effects is skew-normal, and the distribution of the random errors is normal, or verse visa.

There are many different ways to give the definition of skew-normal distributions, because they have so many different characterizations which can derive the same results for a general class of random variables. Two of those definitions are commonly used. One main idea focuses on the conditional representation, which is related to an unobservable random variable X_0 and more appealing conceptual, and the other main idea is to use absolute values to give the marginal representation, which gives

some moments computations and theoretical derivations.

3.2 Definition of the Skew-normal Distribution

We will first introduce the basic definition of the skew-normal distribution. There are many different definitions of skew-normal distributions, and here we will focus on a general definition and an additive definition, which will play an important role in our analysis.

3.2.1 Construction of the Skew-normal Distribution

Symmetric probability density property

Several probability distributions to be inspected in this section can be acquired as special instances of the plan to be presented underneath, which permits us to create an entire arrangement of distributions as a perturbed or regulated form of a symmetric probability density function f_0, which is called the base density. This base is regulated or bothered by an element which can be picked uninhibitedly in light of the fact that it must fulfill extremely basic conditions.

Assume that the density function f_0 is symmetric about a given point x_0 if $f_0(x - x_0) = f_0(x_0 - x)$ for all x, but in the theoretical analysis, we can take $x_0 = 0$ without loss of generality. In the q-dimensional case, the definition of a symmetric density can rather be figured in an assortment of methods. If X, a random variable, is distributed as $-X$ in the same way, then we can say that X is centrally 0 symmetric. In case X is a continuous variable with the density function $f_0(x)$, then central symmetry needs that $f_0(x) = f_0(-x)$ for all $x \in \mathbb{R}^q$. [7]

Proposition 3.1 Suppose f_0 is a probability density function on $\mathbb{R}^q$, $G_0(\cdot)$ is a continuous distribution function on the real line, and $w(\cdot)$ is a real valued function on $\mathbb{R}^q$ such that $f_0(-x) = f_0(x)$, $w(-x) = w(x)$ and $G_0(-y) = 1 - G_0(y)$ for all $x \in \mathbb{R}^q$,

$y \in \mathbb{R}$. [7] Then

$$f(x) = 2f_0(x)G_0\{w(x)\} \tag{3.1}$$

is a density function on $x \in \mathbb{R}^q$.

In the technical proof, take note of that $g(x) = 2[G_0\{w(x)\} - 1/2]f_0(x)$ is an odd function and it is integrable since $\mid g(x) \mid \leq f_0(x)$. Then [7]

$$0 = \int_{R^d} g(x)dx = \int_{R^d} 2f_0(x)G_0\{w(x)\}dx - 1. \tag{3.2}$$

Despite the fact that the proof is sufficient, it does not clarify the part of the different components in a probability perspective. In the proof underneath, we signify by A the set shaped by turning around the indication of all components of A, in the event that A represents an Euclidean space subset. We conclude that A is a symmetric set when $A = -A$.

In informative proof, let Z_0 signify an arbitrary variable with density f_0 and T a variable with distribution G_0, independent of Z_0. To demonstrate that $W = w(Z_0)$ has a distribution symmetric about 0, consider a Borel set A of the real line and present [7]

$$P\{W \in -A\} = P\{-W \in A\} = P\{W(-Z_0) \in A\} = P\{W(Z_0) \in A\} \tag{3.3}$$

considering that Z_0 and $-Z_0$ have the similar distribution. Since T is symmetric about 0, then we presume that

$$\begin{aligned} 1/2 = P\{T \leq W\} &= E_{Z_0}\{P[T \leq w(Z_0) \mid Z_0 = x]\} \\ &= \int_{R^d} G_0\{w(x)\}f_0(x)dx. \end{aligned} \tag{3.4}$$

On setting $G(x) = G_0\{w(x)\}$ in (3.1), we can modify (3.1) as

$$f(x) = 2f_0(x)G(x), \tag{3.5}$$

where

$$G(x) \geq 0, \quad G(x) + G(-x) = 1. \tag{3.6}$$

The other way around, any function G fulfilling (3.6) can be composed in the structure $G_0 w(x)$. Which of the two structures, (3.1) or (3.5), will be utilized relies on upon the specific circumstance. Representation of $G(x)$ in the structure $G_0 w(x)$ is not remarkable since, $G(x) = G_1\{w_1(x)\}$, $w_1(x) = G_1^{-1}[G_0\{w(x)\}]$, for any monotonically increasing distribution function G_1 on the real line fulfilling $G_1(-y) = 1 - G_1(y)$. [7]

Acceptance-rejection method

Also, we will infer the density of skew-normal class by acceptance-rejection method, since it is very hard to find an explicit formula for the cdf of X, defined by $G(x) = P(X \leq x)$. Acceptance-rejection method is one of the most important methods to generate random values and is more efficient than the inverse transform method. [20] The main idea of acceptance-rejection method is to find a function $h(x) = c \cdot g(x)$, where $g(x)$ is the probability density function (pdf) and $H(x)$ is the probability distribution function with support $x \in A$. Here c is a chosen constant that satisfies $sup_x\{g(x)/h(x)\} \leq c$, which means that the ratio $g(x)/h(x)$ is bounded by a constant $c > 0$ and the function $h(x)$ is close to $g(x)$. [5] If we set sample X from the absolutely continuous function H and Y from the density function g, and if X and Y are independent, then we can have

$$P\{X - \lambda Y < 0\} = E_Y(P\{X < \lambda y \mid Y = y\}) = \int H(\lambda y) g(y) dy. \tag{3.7}$$

The principle that acceptance-rejection method works on is that we choose $Z = Y$ when $X < \lambda Y$ or generate another set of X and Y when $X \geq \lambda Y$ until it meets the condition $X < \lambda Y$. [20] More precisely, consider two independent and identically distributed standard normal random variables Y and X. Now define Z to be equal to

Y conditionally on the event $\lambda Y > X$. The resulting distribution of Z is given by [5]

$$\begin{aligned} P(Z \leq z) &= P(Y \leq z \mid \lambda Y > X) \\ &= P(Y \leq z, \lambda Y > X)/P(\lambda Y > X) \\ &= \int_{-\infty}^{z} \int_{-\infty}^{\lambda y} \varphi(y)\varphi(x) dw dy / P(\lambda Y > X) \\ &= \int_{-\infty}^{z} \varphi(y)\Phi(\lambda y) dy / P(\lambda Y > X). \end{aligned} \tag{3.8}$$

Since $P(\lambda Y > X) = P(\lambda Y - X > 0) = 1/2$ (because $\lambda Y - X$ follows a normal distribution with mean 0), we can say that Z has the skew-normal density with respect to z. [5]

3.2.2 Different Univariate Skew-normal Distribution Presentations

As indicated by the symmetric probability density property and acceptance-rejection method, here we present the first definition of the skew-normal density function. [6] The density function is given by

$$f_Y(y) = 2\phi(y)\Phi(\lambda y), \tag{3.9}$$

where Φ is the standard normal cumulative distribution function and ϕ is the probability density function. Furthermore, Y is a skew-normal variable and $\phi(y; \lambda)$ is an alternative density function of Y. We call λ the skewness parameter. [6]

For the most part, we can also express it by the normal distribution with expectation value μ and variance parameter σ^2. [5] Suppose Z is a continuous random variable with the density function (3.9). Then the variable

$$Y = \mu + \sigma Z \tag{3.10}$$

is called a skew-normal (SN) variable with the location parameter μ, the scale parameter σ, and the skewness parameter λ. [5] Its density function at $x \in R$ is given by

$$2\phi(y \mid \mu, \sigma^2)\Phi(\lambda\frac{y-\mu}{\sigma}) = \phi^*(\frac{x-\mu}{\sigma}; \lambda), \tag{3.11}$$

and we denote it by $Y \sim SN(\mu, \sigma^2, \lambda)$.

It could be disentangled as $Y \sim SN_1(\lambda)$ with $\sigma^2 = 1$ and $\mu = 0$. Furthermore, if $\lambda = 0$, the skew-normal distribution becomes a normal distribution. And if $\lambda = +\infty$, the skew-normal distribution becomes a half-normal distribution.

Secondly, we can also give an additive definition of the skew-normal distribution [17]. This definition begins with the i.i.d. standard normal random variables U, V and a constant $\delta \in (-1, 1)$. If we define

$$Z = \delta \mid U \mid +\sqrt{1-\delta^2}V, \tag{3.12}$$

then a simple convolution computation may be used to verify that Z has a skew-normal (λ) distribution with $\lambda = \delta/\sqrt{1-\delta^2}$. In order to proof this, we suppose that $m = \lambda(1+\lambda^2)^{-1/2}$ and $n = (1+\lambda^2)^{-1/2}$. Then we have

$$\begin{aligned} P(Y_\lambda \le y) &= E[P(Y_\lambda \le y \mid\mid U \mid)] \\ &= \int_0^\infty P\{V \le (y-mu)/b\}2\phi(u)du \\ &= 2\int_0^\infty \Phi\{(y-mu)/b\}\phi(u)du, \end{aligned} \tag{3.13}$$

and from the equation $m^2 + n^2 = 1$ we can have the conclusion that

$$\begin{aligned} \frac{d}{dy}P(Y_\lambda \le y) &= 2\phi(y)\int_0^\infty (2\pi n^2)^{-1/2}\exp\{-n^{-2}(u-my)\}du \\ &= 2\phi(y)\left\{1-\Phi\left(-\frac{my}{n}\right)\right\} \\ &= 2\phi(y)\Phi(\lambda y). \end{aligned} \tag{3.14}$$

This is the proof of the additive definition of the skew-normal distribution. [17] We can also rewrite the general idea in terms of the skewness parameter λ that Y_λ defined by

$$Y_\lambda = \frac{\lambda}{(1+\lambda^2)^{1/2}} \mid U \mid + \frac{1}{(1+\lambda^2)^{1/2}} V \tag{3.15}$$

follows the skew-normal distribution $SN_1(\lambda)$ under the assumption that U, V are independent standard normal random variables.

On reflection, the first and second definitions are essentially equivalent. This will become more explicit when we extend univariate cases to multivariate cases in the following section.

3.3 Multivariate Skew-normal Distribution

In this section, we extend the univariate skew-normal distribution to the multivariate skew-normal distribution. We first illustrate the main idea in an additive form. Here, we regard $\mathbf{Y}$ as a q-dimensional random variable with each component being skew-normal.

If the random vector $\mathbf{Y}$ is distributed as skew-normal with the mean vector $\boldsymbol{\mu} \in \mathbb{R}^q$, variance-covariance matrix $\boldsymbol{\Sigma}$ and skewness vector $\boldsymbol{\lambda} \in \mathbb{R}^q$, then its density function is given by [8]

$$f_{\mathbf{Y}}(\mathbf{y}) = 2\phi_q(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \boldsymbol{\Phi}_1(\boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})), \mathbf{y} \in \mathbb{R}^q. \tag{3.16}$$

The simplified form of this distribution is $\mathbf{Y} \sim SN_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. The simplified skew-normal form for $\mu = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I_q}$ is $\mathbf{Y} \sim SN_q(\boldsymbol{\lambda})$.

3.3.1 Additive representation

It is common to define a q-dimensional normal random vector with standardised marginals as $\mathbf{U} = (U_1, \ldots, U_q)^T \sim \mathbf{I}_q$, which is independent with $U_0 \sim N(0,1)$. So we

can obtain

$$\mathbf{Y} = \boldsymbol{\delta}|U_0| + (\mathbf{I_q} - \boldsymbol{\delta}\boldsymbol{\delta}^T)^{1/2}\mathbf{U} \tag{3.17}$$

according to (3.6). Note that $\boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{(1+\boldsymbol{\lambda}^{\mathbf{T}}\boldsymbol{\lambda})^{1/2}}$. Then we can say that $\mathbf{Y} \sim SN_q(\boldsymbol{\lambda})$. Here we will give a simple proof of this additive form. We can see that the skew-normal distribution of the additive form is equal to general form that we introduced earlier. Let $\mathbf{Y} \mid |U_0| = t \sim N_q(\boldsymbol{\delta}t, \mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}^T)$, where $|U_0| \sim HN(0,1)$. Here "HN" means the half-normal distribution, which is fold at mean 0 in the normal distribution. Then by Lemma 3.1 given later, it follows that [3]

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \int_0^\infty \phi_q(\mathbf{y} \mid \boldsymbol{\delta}t, \mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}^T) 2\phi(t) dt \\ &= \int_{\mathbf{0}}^\infty \phi_q(\mathbf{y} \mid \mathbf{0}, \mathbf{I}_q) 2\phi(t \mid \boldsymbol{\delta}^T\mathbf{y}, \mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}^T) dt \\ &= 2\phi_q(\mathbf{y})\Phi_1\left(\frac{\boldsymbol{\delta}^T\mathbf{y}}{\sqrt{\mathbf{I}_q - \boldsymbol{\delta}\boldsymbol{\delta}^T}}\right), \end{aligned} \tag{3.18}$$

where $\boldsymbol{\delta} = \frac{\boldsymbol{\lambda}}{\sqrt{1+\boldsymbol{\lambda}^T\boldsymbol{\lambda}}}$. We need to note that this additive form can separate a skew-normal random variable into two parts, which are one-dimensional standard normal variable and n-dimensional standard normal variable.

Lemma 3.1 Let $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{X} \sim N_q(\boldsymbol{\eta}, \boldsymbol{\Omega})$. Then, [3]

$$\begin{aligned} \phi_p(\mathbf{y} \mid \boldsymbol{\mu} + \mathbf{A}\mathbf{x}, \boldsymbol{\Sigma})\phi_q(\mathbf{x} \mid \boldsymbol{\eta}, \boldsymbol{\Omega}) &= \phi_p(\mathbf{y} \mid \boldsymbol{\mu} + \mathbf{A}\boldsymbol{\eta}, \boldsymbol{\Sigma} + \mathbf{A}\boldsymbol{\Omega}\mathbf{A}^T) \\ &\quad \times \phi_q(\mathbf{x} \mid \boldsymbol{\eta} + \boldsymbol{\Lambda}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta}), \boldsymbol{\Lambda}), \end{aligned} \tag{3.19}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\Omega}^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}$. The density function of $\mathbf{Y}$ is the multivariate normal density

$$f(y_1, \ldots, y_p) = \frac{1}{\sqrt{(2\pi)^p \mid \boldsymbol{\Sigma} \mid^{1/2}}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right\}. \tag{3.20}$$

For simplicity, we ignore the constant part $\frac{1}{\sqrt{(2\pi)^p|\mathbf{\Sigma}|^{1/2}}}$, and let $\mathbf{z} = \mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta}$ and $\mathbf{w} = \mathbf{x} - \boldsymbol{\eta}$.

Then

$$
\begin{aligned}
&\phi_p(\mathbf{y} \mid \boldsymbol{\mu} + \mathbf{A}\mathbf{x}, \mathbf{\Sigma})\phi_q(\mathbf{x} \mid \boldsymbol{\eta}, \mathbf{\Omega}) \\
&= \exp\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta})^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta})\} \times \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\eta})^T\mathbf{\Omega}^{-1}(\mathbf{x} - \boldsymbol{\eta})\} \\
&= \exp\{-\frac{1}{2}(\mathbf{z} - \mathbf{A}\mathbf{w})^T\mathbf{\Sigma}^{-1}(\mathbf{z} - \mathbf{A}\mathbf{w})\} \times \exp\{-\frac{1}{2}\mathbf{w}^T\mathbf{\Omega}^{-1}\mathbf{w}\} \\
&= \exp\left\{-\frac{1}{2}[\mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{z} + \mathbf{w}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{w} + \mathbf{w}^T\mathbf{\Omega}^{-1}\mathbf{w} - \mathbf{w}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z} - \mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{w}]\right\} \\
&= \exp\left\{-\frac{1}{2}[\mathbf{z}^T(\mathbf{\Sigma} + \mathbf{A}\mathbf{\Omega}\mathbf{A}^T)^{-1}\mathbf{z} + \mathbf{w}^T(\mathbf{\Omega}^{-1} + \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})\mathbf{w}]\right\} \\
&\times \exp\left\{-\frac{1}{2}[+\mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z} - \mathbf{w}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z} - \mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{w}]\right\} \\
&= \exp\left\{-\frac{1}{2}[\mathbf{z}^T(\mathbf{\Sigma} + \mathbf{A}\mathbf{\Omega}\mathbf{A}^T)^{-1}\mathbf{z} + \mathbf{w}^T\mathbf{\Lambda}^{-1}\mathbf{w}]\right\} \\
&\times \exp\left\{-\frac{1}{2}[+\mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z} - \mathbf{w}^T\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z} - \mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{w}]\right\} \\
&= \exp\left\{-\frac{1}{2}[\mathbf{z}^T(\mathbf{\Sigma} + \mathbf{A}\mathbf{\Omega}\mathbf{A}^T)^{-1}\mathbf{z} + \mathbf{w}^T\mathbf{\Lambda}^{-1}\mathbf{w}]\right\} \\
&\times \exp\left\{-\frac{1}{2}[+\mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z} - \mathbf{w}^T\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z} - \mathbf{z}^T\mathbf{\Sigma}^{-1}\mathbf{A}\mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{w}]\right\} \\
&= \exp\left\{-\frac{1}{2}[\mathbf{z}^T(\mathbf{\Sigma} + \mathbf{A}\mathbf{\Omega}\mathbf{A}^T)^{-1}\mathbf{z} + (\mathbf{w} - \mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z})^T\mathbf{\Lambda}^{-1}(\mathbf{w} - \mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{z})]\right\} \\
&= \exp\left\{-\frac{1}{2}[(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta})^T(\mathbf{\Sigma} + \mathbf{A}\mathbf{\Omega}\mathbf{A}^T)^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta})]\right\} \\
&\times \exp\left\{-\frac{1}{2}[+(\mathbf{x} - \eta - \mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta}))^T\mathbf{\Lambda}^{-1}(\mathbf{x} - \boldsymbol{\eta} - \mathbf{\Lambda}\mathbf{A}^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu} - \mathbf{A}\boldsymbol{\eta}))]\right\},
\end{aligned}
$$

where $\mathbf{\Lambda} = (\mathbf{\Omega}^{-1} + \mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}$. This completes the proof of the lemma.

3.4 Moment Generating Function of Skew-normal Distribution

In the probability theory, we have another way to get the probability distribution of a certain random variable. That is to say, moment generating function provides the basis of analytical methods to obtain main probability indicators. In addition to univariate distributions, moment generating functions can also be defined for multivariate distributions. There are several relations between properties of the distribution function and behaviors of the moment generating function.

In probability theory and statistics, the moment generating function of a random variable X is defined by

$$M_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx. \tag{3.21}$$

That is, the moment generating function can be regraded as the expectation of the random variable e^{tX}.

3.4.1 Moment generating function of $SN(0, 1, \lambda)$

If U is a random variable which follows $N(0, 1)$, then the moment generating function of U is $M(t) = e^{t^2/2}$. We can derive the moment generating function of skew-normal random variable $Z \sim SN_1(\lambda)$ as

$$M(t) = 2 \exp\left(\frac{t^2}{2}\right) \Phi(\delta t) = 2 \exp\left(\frac{t^2}{2}\right) \Phi\left(\frac{\lambda t}{\sqrt{1+\lambda^2}}\right). \tag{3.22}$$

where $\delta = \frac{\lambda}{\sqrt{1+\lambda^2}}$.

We can use this moment generating function to get the mean and variance of the skew-normal random variable as follows. The first derivative of the moment

generating function is

$$\begin{aligned} M'(t) &= 2t \cdot \exp\left(\frac{t^2}{2}\right)\Phi(\delta t) + 2\exp\left(\frac{t^2}{2}\right)\phi(\delta t)\delta \\ &= 2t \cdot \exp\left(\frac{t^2}{2}\right)\Phi(\delta t) + 2\exp\left(\frac{t^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\delta^2 t^2}{2}\right)\delta \\ &= 2t \cdot \exp\left(\frac{t^2}{2}\right)\Phi(\delta t) + \sqrt{\frac{2}{\pi}}\exp\left(\frac{t^2(1-\delta^2)}{2}\right)\delta. \end{aligned} \tag{3.23}$$

The mean of the skew-normal distribution is obtained as

$$M'(0) = 2 \cdot \frac{1}{\sqrt{2\pi}} = \sqrt{\frac{2}{\pi}}\delta. \tag{3.24}$$

$$\begin{aligned} M''(t) = {} & 2\exp\left(\frac{t^2}{2}\right)\Phi(\delta t) + 2t^2\exp\left(\frac{t^2}{2}\right)\Phi(\delta t) \\ & + 2t\exp\left(\frac{t^2}{2}\right)\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-\delta^2 t^2}{2}\right) + \sqrt{\frac{2}{\pi}}(1-\delta^2)\delta \end{aligned} \tag{3.25}$$

So for the skew-normal random variable $Z \sim SN(0,1,\lambda)$, we have the mean and variance as follows:

$$E(Z) = \sqrt{\tfrac{2}{\pi}}\delta, \quad Var(Z) = 1 - \tfrac{2}{\pi}\delta^2. \tag{3.26}$$

3.4.2 Moment generating function of $SN(\mu, \sigma^2, \lambda)$

Lemma 3.2 While in 1-dimensional case, if $Z \sim N(\xi, \theta^2)$ and $Y \sim \sqrt{\chi^2_m/m}$ is independent of Z, then for any c, [13]

$$E[\Phi(Z + cY)] = P\left\{T(\delta; m) \leq \frac{c}{\sqrt{1+\theta^2}}\right\}, \tag{3.27}$$

where $T(\delta; m)$ is a noncentral t variable with $m$degrees of freedom and noncentral parameter $\delta = -\xi/\sqrt{1+\theta^2}$.

Proof: Let $X \sim N(0,1)$ independently of Y and Z. Then

$$\begin{aligned} E_{Y,Z}[\Phi(Z+cY)] &= k\int_{-\infty}^{\infty}\int_{0}^{\infty}\left[\int_{-\infty}^{z+cy} e^{-\frac{x^2}{2}}dx\right] f(y,z)dydz \\ &= E_{Y,Z}[P_x\{x \le z+cy \mid y,z\}] \\ &= P_{X,Y,Z}\{X \le Z+cY\} \\ &= P_{X,Y,Z}\left\{\frac{X-Z}{Y\sqrt{1+\theta^2}} \le \frac{c}{\sqrt{1+\theta^2}}\right\} \\ &= P\left\{T(\delta;m) \le \frac{c}{\sqrt{1+\theta^2}}\right\}. \end{aligned} \tag{3.28}$$

Set $c = 0$, and note that for all m,

$$\begin{aligned} P\{T(\delta;m) < 0\} &= P\left\{\frac{X+\delta}{Y} \le 0\right\} \\ &= P\{X \le -\delta\} \\ &= \Phi(-\delta) \\ &= \Phi(\xi/\sqrt{1+\theta^2}). \end{aligned} \tag{3.29}$$

From this result, the moment generating function of Y is readily obtained, that is

$$\begin{aligned} M(t) &= E\{\exp(\mu t + \sigma_1 Z t)\} \\ &= 2\exp\left(\mu t + \frac{1}{2}\sigma_1^2 t^2\right)\int_R \phi(z-\sigma_1 t)\Phi(\lambda z)dz \\ &= 2\exp\left(\mu t + \frac{1}{2}\sigma_1^2 t^2\right)\Phi(\delta\sigma_1 t) \end{aligned} \tag{3.30}$$

where $\delta = \delta(\lambda) = \frac{\lambda}{\sqrt{1+\lambda^2}}$. Thus, $\delta \in (-1,1)$.

To compute the moments of $Y \sim SN(\mu, \sigma_1^2, \lambda)$, one route is via the above moment generating function (3.30) or, equivalently but somewhat more conveniently, via the cumulate generating function [7]

$$K(t) = \log M(t) = \mu t + \frac{1}{2}\sigma_1^2 t^2 + \kappa_0 \log\{2\Phi(x)\} \tag{3.31}$$

Let

$$\kappa_0(x) = \log\{2\Phi(x)\}. \tag{3.32}$$

We shall also make use of the derivatives

$$\kappa_r(x) = \frac{d^r}{dx^r}\log\{2\Phi(x)\} = \frac{d^r}{dx^r}\kappa_0(x), \tag{3.33}$$

whose expressions, for the lower orders, are

$$\begin{aligned} \kappa_1(x) &= \phi(x)/\Phi(x) \\ \kappa_2(x) &= -\kappa_1(x)\{x + \kappa_1(x)\} = -\kappa_1(x)^2 - x\kappa_1(x). \end{aligned} \tag{3.34}$$

All $\kappa_r(x)$ for $r > 1$ can be written as functions of $\kappa_1(x)$ and powers of x. For later use, notice that

$$\kappa_1(x) > 0, \quad x + \kappa_1(x) > 0, \quad \kappa_2(x) < 0. \tag{3.35}$$

Then we can have

$$E(Y) = \mu + \sigma_1\mu_z, \tag{3.36}$$

$$Var(Y) = (\sigma_1\sigma_z)^2, \tag{3.37}$$

where

$$\mu_z = E(Z) = b\delta, \quad \sigma_z^2 = var(Z) = 1 - \mu_z^2 = 1 - b^2\delta^2. \tag{3.38}$$

Lemma 3.3 If $\mathbf{A}$ is a symmetric positive definite $d \times d$ matrix, $\mathbf{a}$ and $\mathbf{c}$ are d-dimensional vectors and c_0 is a scalar [7], then

$$\begin{aligned} I &= \int_{\mathbf{R}^q} \frac{1}{(2\pi)^{d/2}\det|\mathbf{A}|^{1/2}} \exp\left[\frac{1}{2}(\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x} - 2\mathbf{a}^T\mathbf{x})\right]\Phi(c_0 + c^T\mathbf{x})d\mathbf{x} \\ &= \exp\left(\frac{1}{2}\mathbf{a}^T\mathbf{A}\mathbf{a}\right)\Phi\left(\frac{c_0 + \mathbf{c}^T\mathbf{A}\mathbf{a}}{\sqrt{1 + \mathbf{c}^T\mathbf{A}\mathbf{c}}}\right) \end{aligned} \tag{3.39}$$

Proof: In the integrand of I, rewrite $\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x} - 2\mathbf{a}^T\mathbf{x}$ as $(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \boldsymbol{\mu}^T\mathbf{A}^{-1}\boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbf{A}\mathbf{a}$, so that

$$I = \exp\left(\frac{1}{2}\mathbf{a}^T\mathbf{A}\mathbf{a}\right)\int_{\mathbf{R}^q}\phi(\mathbf{y};\mathbf{A})\Phi\{c_0 + \mathbf{c}^T(\mathbf{y}+\boldsymbol{\mu})\}d\mathbf{y} \tag{3.40}$$

after a change of variable.

An extension of above Lemma 3.3 for skew-normal variables can be obtained as follows. [7]

Lemma 3.4 If $Z \sim SN(0,1,\lambda)$ and $U \sim N(0,1)$, then

$$E\{\Phi(hZ+k)\} = \Phi\left(\frac{k}{\sqrt{1+h^2}}; -\frac{h\lambda}{\sqrt{1+h^2+\lambda^2}}\right), \tag{3.41}$$

$$E\{\Phi(hU = K;\lambda)\} = \Phi\left(\frac{k}{\sqrt{1+h^2}}; \frac{\lambda}{\sqrt{1+h^2(1+\lambda^2)}}\right). \tag{3.42}$$

Proof: If $Z \sim SN(0,1,\lambda)$ and $U \sim N(0,1)$ are independent variables, then

$$\begin{aligned} E\{\Phi(hZ+k)\} &= E\{P\{U \le hz+k \mid Z = z\}\} \\ &= P\{U - hZ \le k\} \end{aligned}$$

and, by applying the lemma presented before to the distribution of $\mathbf{U} - \mathbf{h}\mathbf{Z}$, we arrive at the first statement below; the second one is obtained in a similar way. We should note that $\mathbf{h}^T\mathbf{U} \sim N(\mathbf{0}, \mathbf{h}^T\boldsymbol{\Sigma}\mathbf{h})$ if $\mathbf{U} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$. The subsequent statement illustrates the technique of completing the square for a skew-normal type of integrand.

In order to do some preparation for the following derivation, here we give the q-dimensional extension of the above lemmas.

Lemma 3.5 Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then for any fixed q-dimensional vector $\mathbf{a}$ and $q \times n$ matrix $\mathbf{B}$, [3]

$$E[\Phi_q(\mathbf{a} + \mathbf{B}\mathbf{Y} \mid \boldsymbol{\eta}, \boldsymbol{\Omega})] = \Phi_q(\mathbf{a} \mid \boldsymbol{\eta} - \mathbf{B}\boldsymbol{\mu}, \boldsymbol{\Omega} + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T). \tag{3.43}$$

Proof:

$$\begin{aligned} E[\Phi_k(\mathbf{a} + \mathbf{B}\mathbf{Y} \mid \boldsymbol{\eta}, \boldsymbol{\Omega})] &= E[\Phi_k(\mathbf{a} \mid \boldsymbol{\eta} - \mathbf{B}\mathbf{Y}, \boldsymbol{\Omega})] \\ &= E[P(\mathbf{U} \leq \mathbf{a} \mid \mathbf{Y})] \\ &= P(\mathbf{U} \leq \mathbf{a}). \end{aligned} \tag{3.44}$$

Thus,

$$\begin{aligned} Var(\mathbf{U}) &= E[Var(\mathbf{U} \mid \mathbf{Y})] + Var[E(\mathbf{U} \mid \mathbf{Y})] \\ &= E[\boldsymbol{\Omega}] + Var(\boldsymbol{\eta} - \mathbf{B}\mathbf{y}) \\ &= \boldsymbol{\Omega} + \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T, \end{aligned} \tag{3.45}$$

where $\mathbf{U} \mid \mathbf{Y} = \mathbf{y} \sim N_q(\boldsymbol{\eta} - \mathbf{B}\mathbf{y}, \boldsymbol{\Omega})$, so that $\mathbf{U} \sim \mathbf{N}_q(\boldsymbol{\eta} - \mathbf{B}\boldsymbol{\mu}, \boldsymbol{\Omega} + \mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T)$.

To compute the moment generating function $M(t)$ of $\mathbf{Y} \sim SN_q(\boldsymbol{\mu}, \boldsymbol{\Omega}, \boldsymbol{\lambda})$, we write $\mathbf{Y} = \boldsymbol{\mu} + \sigma\mathbf{Z}$, where $\mathbf{Z} \sim SN_q(\boldsymbol{\mu}, \bar{\boldsymbol{\Omega}}, \boldsymbol{\lambda})$. Then, using Lemma 5.3, we obtain [1]

$$\begin{aligned} M(\mathbf{t}) &= \exp(\mathbf{t}^T\boldsymbol{\mu}) \int_{R^q} 2\exp(\mathbf{t}^T\boldsymbol{\mu}\mathbf{z})\phi_q(\mathbf{z}; \bar{\boldsymbol{\Omega}})\Phi(\boldsymbol{\lambda}^T\mathbf{z})d\mathbf{z} \\ &= 2\exp\left(\mathbf{t}^T\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Omega}\mathbf{t}\right)\Phi(\sigma\boldsymbol{\delta}^T\mathbf{t}), \end{aligned} \tag{3.46}$$

where

$$\boldsymbol{\delta} = (1 + \boldsymbol{\lambda}^T\bar{\boldsymbol{\Omega}}\boldsymbol{\lambda})^{-1/2}\bar{\boldsymbol{\Omega}}\boldsymbol{\lambda}. \tag{3.47}$$

For later use, we write down the inverse relationship:

$$\boldsymbol{\lambda} = \frac{\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}}{\sqrt{1 - \boldsymbol{\delta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\delta}}} \tag{3.48}$$

To summarize, according to the moment generating function of skew-normal distribution which is derived before, we can have the following result.

Proposition 3.2 Let $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{W}$, where $\mathbf{W} \sim SN_n(\boldsymbol{\lambda})$. Then $\mathbf{Y} \sim SN_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$. Moreover,

$$\begin{aligned} E[\mathbf{Y}] &= \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}, \\ Var[\mathbf{Y}] &= \boldsymbol{\Sigma} - \frac{2}{\pi}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\delta}\boldsymbol{\delta}^T\boldsymbol{\Sigma}^{1/2}, \end{aligned} \tag{3.49}$$

as showing in moment generating function in previous lemmas. [3]

Lemma 3.6 Considering $\mathbf{b}_i \sim SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b)$ and $\boldsymbol{\epsilon}_i \sim SN(\mathbf{0}, \boldsymbol{\psi}_i, \boldsymbol{\lambda}_{ej})$ which are independent, we can rewrite the Lemma 3.1 as follows [3]

$$\phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\psi})\phi_q(\mathbf{b} \mid \mathbf{0}, \mathbf{D}) = \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})\phi_q(\mathbf{b} \mid \boldsymbol{\mu}_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Lambda}) \tag{3.50}$$

and

$$\Phi_1(\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}))\Phi_1(\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\mathbf{b}) = \Phi_2(-\boldsymbol{\Gamma}\mathbf{b} \mid -\boldsymbol{\mu}_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{I}_2), \tag{3.51}$$

where $\boldsymbol{\mu}_1 = \boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}$, $\boldsymbol{\Sigma} = \boldsymbol{\psi} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$, and $\boldsymbol{\Lambda} = (\mathbf{D}^{-1} + \mathbf{Z}^T\boldsymbol{\psi}^{-1}\mathbf{Z})^{-1}$. Note that

$$\boldsymbol{\mu}_2 = \begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2} \\ 0 \end{pmatrix},$$

and

$$\boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z} \\ -\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2} \end{pmatrix}.$$

Proof: According to Lemma 3.1,

$$\begin{aligned}
&\phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\psi})\phi_q(\mathbf{b} \mid \mathbf{0}, \mathbf{D}) \\
&= \phi_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\psi} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T)\phi_q(\mathbf{b} \mid \mathbf{0} + (\mathbf{D}^{-1} + \mathbf{Z}^T\boldsymbol{\psi}^{-1}\mathbf{Z})^{-1}\mathbf{Z}^T\boldsymbol{\psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), (\mathbf{D}^{-1} + \mathbf{Z}^T\boldsymbol{\psi}^{-1}\mathbf{Z})^{-1}) \\
&= \phi_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})\phi_q(\mathbf{b} \mid \boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Lambda}) \\
&= \phi_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})\phi_q(\mathbf{b} \mid \boldsymbol{\mu}_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Lambda}),
\end{aligned} \tag{3.52}$$

where $\boldsymbol{\Sigma} = \boldsymbol{\psi} + \mathbf{Z}\mathbf{D}\mathbf{Z}^T$, $\boldsymbol{\Lambda} = (\mathbf{D}^{-1} + \mathbf{Z}^T\boldsymbol{\psi}^{-1}\mathbf{Z})^{-1}$ and $\boldsymbol{\mu}_1 = \boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}$.

Suppose

$$\mathbf{W} = \begin{pmatrix} U \\ V \end{pmatrix} \sim \mathcal{N}_2 \left\{ \mathbf{0} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right\}$$

and

$$\begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b}) \\ \boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\mathbf{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ 0 \end{pmatrix} + \begin{pmatrix} -\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z}\mathbf{b} \\ \boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\mathbf{b} \end{pmatrix}$$

$$\mathbf{T} = \mathbf{W} + \boldsymbol{\Gamma}\mathbf{b} = \begin{pmatrix} U \\ V \end{pmatrix} + \begin{pmatrix} -\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z} \\ -\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2} \end{pmatrix} \mathbf{b} \sim N_2(\boldsymbol{\Gamma}\mathbf{b}, \mathbf{I}_2).$$

Then

$$\begin{aligned}
P\left(\mathbf{T} \leq \begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\mathbf{y} - X\boldsymbol{\beta}) \\ 0 \end{pmatrix}\right) &= \Phi_2\left(\begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\mathbf{y} - X\boldsymbol{\beta}) \\ 0 \end{pmatrix} \mid \boldsymbol{\Gamma}\mathbf{b}, \mathbf{I}_2\right) \\
&= \Phi_2(-\boldsymbol{\Gamma}\mathbf{b} \mid -\boldsymbol{\mu}_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{I}_2)
\end{aligned}$$

where

$$\boldsymbol{\mu}_2 = \begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2} \\ 0 \end{pmatrix}.$$

Theorem 1 Let $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i$, where $\mathbf{b}_i \sim SN_q(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b)$ and $\boldsymbol{\epsilon}_i \sim SN(\mathbf{0}, \boldsymbol{\psi}_i, \boldsymbol{\lambda}_{ej})$ are independent. Then, the marginal distribution of $\mathbf{Y}_i$ is given by [4]

$$\begin{aligned} f_{\mathbf{Y}_i}(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) &= 2^2 \phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i) \\ &\quad \times \left((\boldsymbol{\mu}_{2j} - \boldsymbol{\Gamma}_i\boldsymbol{\mu}_{1j})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \mid \mathbf{0}, \mathbf{I}_2 + \boldsymbol{\Gamma}_i\boldsymbol{\Lambda}_i\boldsymbol{\Gamma}_i^T\right) \end{aligned} \tag{3.53}$$

where $\boldsymbol{\mu}_{1j} = \boldsymbol{\Lambda}_i\mathbf{Z}_i^T\boldsymbol{\psi}_i^{-1}$, $\boldsymbol{\Sigma}_i = \boldsymbol{\psi}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$, and $\boldsymbol{\Lambda}_i = (\mathbf{D}^{-1} + \mathbf{Z}_i^T\boldsymbol{\psi}_i^{-1}\mathbf{Z}_i)^{-1}$. Note that

$$\boldsymbol{\mu}_{2j} = \begin{pmatrix} \boldsymbol{\lambda}_{ej}^T\boldsymbol{\psi}_i^{-1/2} \\ 0 \end{pmatrix}$$

and

$$\boldsymbol{\Gamma}_i = \begin{pmatrix} \boldsymbol{\lambda}_{ej}^T\boldsymbol{\psi}_i^{-1/2}\mathbf{Z}_i \\ -\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2} \end{pmatrix}$$

Proof: Using Lemmas 3.1, 3.4 and 3.5, we can have that

$$\begin{aligned} & f_{\mathbf{Y}}(\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \int_{\mathbb{R}^q} f(\mathbf{y} \mid \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_e) f(\mathbf{b} \mid \boldsymbol{\alpha}, \boldsymbol{\lambda}_e) d\mathbf{b} \\ &= \int_{\mathbb{R}^q} 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\psi}) \Phi_1(\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})) \phi_q(\mathbf{b} \mid \mathbf{0}, \mathbf{D}) \Phi_1(\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\mathbf{b}) d\mathbf{b} \\ &= \int_{\mathbb{R}^q} 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\psi}) \phi_q(\mathbf{b} \mid \mathbf{0}, \mathbf{D}) \Phi_1(\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{b})) \Phi_1(\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\mathbf{b}) d\mathbf{b} \\ &= \int_{\mathbb{R}^q} 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \boldsymbol{\psi}) \phi_q(\mathbf{b} \mid \mathbf{0}, \mathbf{D}) \Phi_2(-\boldsymbol{\Gamma}\mathbf{b} \mid -\boldsymbol{\mu}_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{I}_2) d\mathbf{b} \\ &= \int_{\mathbb{R}^q} 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \phi_q(\mathbf{b} \mid \boldsymbol{\mu}_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Lambda}) \Phi_2(-\boldsymbol{\Gamma}\mathbf{b} \mid -\boldsymbol{\mu}_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{I}_2) d\mathbf{b} \\ &= 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \int_{\mathbb{R}^q} \Phi_2(-\boldsymbol{\Gamma}\mathbf{b} \mid -\boldsymbol{\mu}_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{I}_2) \phi_q(\mathbf{b} \mid \boldsymbol{\mu}_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Lambda}) d\mathbf{b} \\ &= 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) E[\Phi_2(-\boldsymbol{\Gamma}\mathbf{W} \mid -\boldsymbol{\mu}_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \mathbf{I}_2)] \\ &= 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \Phi_2[\mathbf{0} \mid -\boldsymbol{\mu}_2(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\Gamma}(\boldsymbol{\mu}_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})), \mathbf{I}_2 + \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T] \\ &= 2^2 \phi_n(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \Phi_2[(\boldsymbol{\mu}_2 - \boldsymbol{\Gamma}\boldsymbol{\mu}_1)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \mid \mathbf{0}, \mathbf{I}_2 - \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T], \end{aligned} \tag{3.54}$$

where $\mathbf{W} \sim N_q(\boldsymbol{\mu}_1(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{\Lambda})$.

The result presented in Theorem 1 is important because it avoids using complex numerical techniques, given that it allows a closed form for the marginal distribution of $\mathbf{Y}_i$, $i = 1, \ldots, m$, facilitating straightforward implementation of inferences with standard optimization routines. Thus, denoting the log-likelihood function by $\ell(\theta, \lambda)$, it can be written as [4]

$$\begin{aligned}\ell(\boldsymbol{\theta}, \boldsymbol{\lambda}) \propto & -\frac{1}{2}\sum_{i=1}^{m}\log \mid \boldsymbol{\Sigma}_i \mid -\frac{1}{2}\sum_{i=1}^{m}\{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\} \\ & \sum_{i=1}^{m}\log\Phi_2\left((\boldsymbol{\mu}_{2j} - \boldsymbol{\Gamma}_i\boldsymbol{\mu}_{1j})(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \mid \mathbf{0}, \mathbf{I}_2 + \boldsymbol{\Gamma}_i\boldsymbol{\Lambda}_i\boldsymbol{\Gamma}_i^T\right)\end{aligned} \tag{3.55}$$

where $\boldsymbol{\mu}_{1j}$, $\boldsymbol{\mu}_{2j}$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Lambda}_i$ are as defined as before.

We should note that no explicit solution is available for the maximization problem so that the likelihood function has to be maximized numerically. Some special cases may be of interest. For instance, the situation where $\boldsymbol{\lambda}_{e1} = \ldots = \boldsymbol{\lambda}_{em} = \mathbf{0}$ or $\boldsymbol{\lambda}_b = \mathbf{0}$, which are special cases of the above general situation.

Corollary 3.1 Under the conditions of Theorem 1, it follows that [3]

(i) if $\boldsymbol{\lambda}_{ej} = \mathbf{0}$, $i = 1, \ldots, m$, then

$$f_{\mathbf{Y}_i}(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}_b) = 2\phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)\Phi_1(\bar{\boldsymbol{\lambda}}_{b_i}^T\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})), \tag{3.56}$$

that is,

$$\mathbf{Y}_i \sim SN_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \bar{\boldsymbol{\lambda}}_{b_i}),$$

with

$$\bar{\boldsymbol{\lambda}}_{b_i} = \frac{\boldsymbol{\Sigma}_i^{-1/2}\mathbf{Z}_i\mathbf{D}^{1/2}\boldsymbol{\lambda}_b}{\sqrt{1 + \boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}_i\mathbf{D}^{-1/2}\boldsymbol{\lambda}_b}}.$$

(ii) if $\boldsymbol{\lambda}_b = \mathbf{0}$, then

$$f_{\mathbf{Y}_i}(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}_{e_i}) = 2\phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)\Phi_1(\bar{\boldsymbol{\lambda}}_{e_i}^T\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})), \tag{3.57}$$

that is,

$$\mathbf{Y}_i \sim SN_{n_i}(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i, \bar{\boldsymbol{\lambda}}_{e_i}),$$

with

$$\bar{\boldsymbol{\lambda}}_{e_i} = \frac{\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\psi}_i^{1/2}\boldsymbol{\lambda}_{e_i}}{\sqrt{1+\boldsymbol{\lambda}_{e_i}^T\boldsymbol{\psi}_i^{-1/2}\mathbf{Z}_i\boldsymbol{\Lambda}_i\mathbf{Z}_i^T\boldsymbol{\psi}_i^{-1/2}\boldsymbol{\lambda}_{e_i}}}.$$

Although simpler, the log-likelihood functions that follow from corollary 3.1 must also be maximized numerically. The asymptotic covariance matrix of parameters estimators (MLE) can be estimated by using the Hessian matrix, which can also be computed numerically by using the program R, for example. In the Chapter 5, we will present an EM-type algorithm for computing the MLE of densities obtained in Corollary 3.1.

Proof of Corollary 3.1: we can write

$$(\boldsymbol{\mu}_2 - \boldsymbol{\Gamma}\boldsymbol{\mu}_1)\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}(\boldsymbol{\psi} - \mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T)\boldsymbol{\psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{pmatrix}$$

and

$$\mathbf{I}_2 + \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T - \begin{pmatrix} 1+\boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1/2}\boldsymbol{\lambda}_e & -\boldsymbol{\lambda}_e^T\psi^{-1/2}\mathbf{Z}\boldsymbol{\Lambda}\mathbf{D}^{-1/2}\boldsymbol{\lambda}_b \\ -\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}\boldsymbol{\lambda}_e & 1-\boldsymbol{\lambda}_b^T D^{-1/2}\boldsymbol{\Lambda}\mathbf{D}^T\boldsymbol{\lambda}_b \end{pmatrix},$$

which is a diagonal matrix when $\boldsymbol{\lambda}_e = \mathbf{0}$ or $\boldsymbol{\lambda}_b = \mathbf{0}$. Hence, for $\boldsymbol{\lambda}_e = \mathbf{0}$, the asymmetric part of Theorem 1 can be computed as

$$\begin{aligned}\Phi_2\left(\boldsymbol{\mu}_2 - \boldsymbol{\Gamma}\boldsymbol{\mu}_1|\mathbf{0},\ \mathbf{I}_2 + \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T\right) &= \Phi_2\left(\left(\mathbf{I}_2 + \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T\right)^{-1/2}(\boldsymbol{\mu}_2 - \boldsymbol{\Gamma}\boldsymbol{\mu}_1)\right) \\ &= \frac{1}{2}\Phi_1\left(\frac{\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sqrt{1+\boldsymbol{\lambda}_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}\mathbf{D}^{-1/2}\boldsymbol{\lambda}_b}}\right),\end{aligned}$$

where some algebraic manipulations $\boldsymbol{\Lambda}\mathbf{Z}^T = \mathbf{D}\mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}$. Similarly, for $\boldsymbol{\lambda}_b = \mathbf{0}$, we have

$$\Phi_2\left(\boldsymbol{\mu}_2 - \boldsymbol{\Gamma}\boldsymbol{\mu}_1|\mathbf{0},\ \mathbf{I}_2 + \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T\right) = \frac{1}{2}\Phi_1\left(\frac{\boldsymbol{\lambda}_c^T\boldsymbol{\psi}^{-1/2}(\boldsymbol{\psi} - \mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T)\boldsymbol{\psi}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\sqrt{1 + \boldsymbol{\lambda}_e^T\boldsymbol{\psi}^{-1/2}\mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T\boldsymbol{\psi}^{-1/2}\boldsymbol{\lambda}_e}}\right),$$

noting that $\boldsymbol{\psi} - \mathbf{Z}\boldsymbol{\Lambda}\mathbf{Z}^T = \boldsymbol{\psi}\boldsymbol{\Sigma}^{-1}\boldsymbol{\psi}$.

Chapter 4

EM Algorithm and Newton-Raphson Method

4.1 Overview of EM Algorithm

A maximization method for the likelihood or log-likelihood may posture issues at times. When we work on advanced studies to deal with complex computational problems, the approach appears not very robust for starting values, that is, if we do not set proper starting values at the beginning, the maximization method may not converge to a local maximal value or may use a really long time. In many simulation studies, the EM method is found to be more robust to some extent than the direct maximization approach.

The fundamental thought of the EM algorithm is to connect the given incomplete-data probability with a complete-data problem. For example, the complete-data problem may produce a closed form roots to the maximum likelihood estimate (MLE) and we can achieve that by a software package of MLE computation. The approach of EM algorithm concentrates on establishing a relationship between the likelihoods of these two problems, and exploiting the less MLE computation of the incomplete-data issue in the M-step of the iterative procedure.

4.2 Maximum likelihood estimation based on different algorithms

Consider multivariate data that are observed $n \times d$ matrix $\mathbf{x} = (x_1^T, x_2^T, \ldots, x_n^T)^T$ with the density function $f(\mathbf{x}; \mathbf{\Psi})$, where $\mathbf{\Psi} = (\Psi_1, \ldots \Psi_d)^T$ is the d-dimensional vector of unknown parameters. The objective of the analysis is to estimate the parameter vector $\mathbf{\Psi}$ that defines the density function f. The likelihood function formed by the observed data $\mathbf{x}$ is given by [22]

$$L(\mathbf{x}; \mathbf{\Psi}) = \prod_{i=1}^{n} f(\mathbf{x}; \mathbf{\Psi}).$$

An estimate $\hat{\mathbf{\Psi}}$ of $\mathbf{\Psi}$ can be obtained as a solution to the likelihood score equations

$$\partial L(\mathbf{\Psi})/\partial \mathbf{\Psi} = 0.$$

Equivalently, we can use the log-likelihood and solve

$$\partial \log L(\mathbf{\Psi})/\partial \mathbf{\Psi} = 0$$

with respect to $\mathbf{\Psi}$ for the ML estimator of $\mathbf{\Psi}$. Briefly, the aim of the ML estimation is to determine estimates $\hat{\mathbf{\Psi}}$, which are defined as the roots of the above likelihood equations that are consistent and asymptotically efficient. We can find such a sequence of roots when we ensure suitable regularity conditions. [24] With the probability of likelihood or log-likelihood tending to 1, these roots correspond to local maxima in the interior of the parameter space.

From the log-likelihood, the observed Fisher information matrix is observed as

$$\mathbf{I}_0(\mathbf{\Psi}; \mathbf{x}) = -\partial^2 \log L(\mathbf{\Psi})/\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T,$$

which is the negative of the second-order partial derivatives of the log-likelihood function with respect to the elements of parameters set $\mathbf{\Psi}$. The expected Fisher information matrix, which is also called the Fisher matrix, $\mathbf{I}(\mathbf{\Psi})$ is given by

$$\begin{aligned}\mathbf{I}(\mathbf{\Psi}) &= E_{\mathbf{\Psi}}\{S(\mathbf{X};\mathbf{\Psi})S^T(\mathbf{X};\mathbf{\Psi})\} \\ &= -E_{\mathbf{\Psi}}\{I_0(\mathbf{\Psi};\mathbf{X})\},\end{aligned}$$

where $S(\mathbf{X};\mathbf{\Psi}) = \partial \log L(\mathbf{\Psi})/\partial\mathbf{\Psi}$ is the gradient vector of the log-likelihood function, which is also called the score statistic. The variance-covariance matrix of the maximum likelihood estimator $\hat{\mathbf{\Psi}}$ is equal to the inverse of the expected information matrix $\mathbf{I}(\mathbf{\Psi})$, which can be approximated by substituting $\mathbf{I}(\mathbf{\Psi})$ by $\mathbf{I}(\hat{\mathbf{\Psi}})$. We can also get the standard error of $\hat{\mathbf{\Psi}}_i$, which is given by $SE(\hat{\mathbf{\Psi}}_i) \approx (\mathbf{I}^{-1}(\hat{\mathbf{\Psi}}))_{ii}^{1/2}$. We often use the observed information matrix to approximate the standard errors, since it is more convenient to use as compared to the expected information matrix. Since the log-likelihood function cannot be maximized analytically in real cases, and we need to compute the maximum likelihood estimate of $\mathbf{\Psi}$ iteratively by using the Newton-Raphson maximization procedure or another algorithm. And we should note that if the total number of parameters, p, is really large in the model, the efficiency of the algorithm will decrease rapidly. [23] Here we give an introduction to the EM algorithm as an alternative method to estimate model parameters and compare it to the Newton-Raphson method.

4.3 EM Algorithm

Before introducing the EM algorithm, we first discuss the theory of DLR, which focuses on the relationship between complete data $\mathbf{x}$ and its density function, and incomplete data $\mathbf{y}$ and its density function. The DLR theory is proposed by Dempster, Laird, and Rubin (1980), who showed that the Iteratively Reweighted Least Squares procedure is an EM algorithm under distributional assumptions. [11]

4.3.1 Theory of DLR

Suppose $\mathbf{x}$ represents complete data and $f(\mathbf{x} \mid \mathbf{\Psi})$ its probability density function. In addition, we assume $\mathbf{y}$ to be incomplete data and $g(\mathbf{y} \mid \mathbf{\Psi})$ to be a probability density function of $\mathbf{y}$. We consider two sample spaces $\mathbf{\Omega_X}$ and $\mathbf{\Omega_Y}$, a sample space for complete data and a sample space for incomplete data and there is a mapping of incomplete data $\mathbf{y} \to \mathbf{y}(\mathbf{x})$ from $\mathbf{\Omega_X}$ to $\mathbf{\Omega_Y}$. Then the probability density function of $\mathbf{y}$, $g(\mathbf{y} \mid \mathbf{\Psi})$, is given by

$$g(\mathbf{y} \mid \mathbf{\Psi}) = \int_{\mathbf{\Omega_Y(y)}} f(\mathbf{x} \mid \mathbf{\Psi}) d\mathbf{x}, \tag{4.1}$$

where $\mathbf{\Omega_{Y(y)}}$ is a subsample space of $\mathbf{\Omega_X}$, which is determined by the equation $\mathbf{y} = \mathbf{y}(\mathbf{x})$. [11] For incomplete data problem, DLR theory assumes that parameters that we want to estimate are independent of the process of generating missing data, that is, missing data are missing at random.

Here we use $l_c(\mathbf{\Psi}) = \log f(\mathbf{x} \mid \mathbf{\Psi})$ to represent the log-likelihood function, which is based on the complete data, and we also use the $l(\mathbf{\Psi}) = \log g(\mathbf{y} \mid \mathbf{\Psi})$ to represent the log-likelihood function, which is based on the incomplete data. The intent of the EM algorithm is to find the maximum likelihood estimator of $\hat{\mathbf{\Psi}}$, which is the point of attaining the maximum of $l(\hat{\mathbf{\Psi}})$.

The EM algorithm approach is indirectly the problem of maximizing the log-likelihood $l(\mathbf{\Psi})$ based on incomplete data by proceeding iteratively in terms of the log-likelihood based on the complete data, $l_c(\mathbf{\Psi})$. Because it is unobservable, it is replaced by the conditional expectation given the observation $\mathbf{y}$ and temporary values of parameters $\tilde{\mathbf{\Psi}}$. We can express the main idea in the following formulae, that is

$$\mathbf{\Psi}^{(k+1)} = \arg\max_{\mathbf{\Psi} \in \mathbf{\Theta}} E[l_c(\mathbf{\Psi}) \mid \mathbf{y}, \mathbf{\Psi}^{(k)}] \tag{4.2}$$

We can divide the above formulae into the E-step and the M-step as follows:

We first set $\mathbf{\Psi}^{(0)}$ to be some initial values for $\mathbf{\Psi}$. On the first iteration, the E-step requires the calculation of

$$Q(\mathbf{\Psi}; \mathbf{\Psi}^{(0)}) = E_{\mathbf{\Psi}^{(0)}}\{l_c(\mathbf{\Psi}) \mid \mathbf{y}\}. \tag{4.3}$$

The M-step for the first iteration requires maximization of $Q(\mathbf{\Psi}, \mathbf{\Psi}^{(0)})$ with respect to $\mathbf{\Psi}$ over the parameter space $\mathbf{\Omega}$. That is, we choose $\mathbf{\Psi}^{(1)}$ which can satisfy

$$Q(\mathbf{\Psi}^{(1)}; \mathbf{\Psi}^{(0)}) \geq Q(\mathbf{\Psi}; \mathbf{\Psi}^{(0)}), \tag{4.4}$$

for all $\mathbf{\Psi} \in \mathbf{\Omega}$. The E-step and M-step are carried out repeatedly until convergence. For the $(k+1)$th iteration, the E-step and M-step are defined as follows:

E-step: Calculate the conditional expectation of complete data log-likelihood given the observation $\mathbf{y}$ and the kth temporary values of parameter $\mathbf{\Psi}^{(k)}$:

$$Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)}) = E[l_c(\mathbf{\Psi}) \mid \mathbf{y}, \mathbf{\Psi}^{(k)}]. \tag{4.5}$$

M-step: Then find $\mathbf{\Psi}^{(k+1)}$ by maximizing $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)})$, calculated in the E-step:

$$Q(\mathbf{\Psi}^{(k+1)}; \mathbf{\Psi}^{(k)}) \geq Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)}). \tag{4.6}$$

The E-step and M-step are repeated by the kth temporary values of parameter $\mathbf{\Psi}^{(k)}$ until they converge to a specified value. That is, the E-steps and M-steps are alternated repeatedly until the difference $l(\mathbf{\Psi}^{(k+1)}) - l(\mathbf{\Psi}^{(k)})$ changes by an arbitrarily small amount in the cases of convergence of the sequence of likelihood values $l(\mathbf{\Psi}^{(k)})$. The DLR theory shows us that the incomplete-data likelihood function $l(\mathbf{\Psi})$ is not decreased after an EM iteration, which means that $l(\mathbf{\Psi}^{(k+1)}) \geq l(\mathbf{\Psi}^{(k)})$.

Suppose that we have already selected proper starting values and then we can get the maximum likelihood estimates of $\mathbf{\Psi}$. If $f(\mathbf{x} \mid \mathbf{\Psi})$ has the regular exponential

family form, we can simplify a very simple characterization of the EM algorithm as follows:

E-step: We calculate the conditional expectation of the complete data in the kth step, which is represented by sufficient statistics $\mathbf{t}$ given observation data $\mathbf{y}$ and $\mathbf{\Psi}^{(k)}$,

$$\mathbf{t}^{(k+1)} = E[\mathbf{t}(\mathbf{x}) \mid \mathbf{y}, \mathbf{\Psi}^{(k)}]. \tag{4.7}$$

M-step: Then we can get $\mathbf{\Psi}^{(k+1)}$ to solve the following equations:

$$E[\mathbf{t}(\mathbf{x})\mathbf{\Psi}] = \mathbf{t}^{(k+1)} \tag{4.8}$$

The above equation is the general form of the likelihood equations for maximum likelihood estimation given data from a regular exponential family, including but not limited to multivariate normal distribution. The score function constructs the likelihood equation for $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ based on the marginal likelihood of all observed data. However, the EM algorithm is a way to maximize the marginal likelihood relating to observed data without using this equation.

Generally speaking, the EM algorithm divides the process of solving above equation into two steps. The first step is to find the conditional expected values on the left side of the equation. The second is to solve the equation in concrete terms using the conditional expected value that has been found. [11]

The drawback of the EM algorithm is that it usually requires a larger number of iterations to reach a convergence. However, it can avoid so-called inverse matrix calculations. Also, programming is easy because a simple algorithm normally applied to complete data can be used in the maximization step.

4.4 ECM algorithm

4.4.1 Introduction

The ECM algorithm (expectation-conditional maximization algorithm) is one of the useful extensions of the EM algorithm, which is proposed by Meng and Rubin (1993) [21]. In these situations, the M-step of the maximization procedure is quite simple since we have conditional function of the parameters estimators. The expectation-conditional maximization algorithm (ECM algorithm), therefore, replaces the M-step of the EM algorithm by a number of computationally simpler conditional maximization (CM) steps. As a result, it usually converges more slowly than the EM algorithm, but it can work faster in total computation time. More importantly, the expectation-conditional maximization algorithm (ECM algorithm) preserves the appealing convergence properties of the EM algorithm, such as its monotone convergence.

As we discussed in the previous section, one of the major reasons for the popularity of the EM algorithm is that the M-step involves only complete-data ML estimation, which is often computationally simple. But if the complete-data ML estimation is rather complicated, then the EM algorithm is less attractive because the M-step is computationally unattractive. In many cases, however, complete-data ML estimation is relatively simple if maximization is undertaken conditional on some of the parameters (or some functions of the parameters). To this end, Meng and Rubin(1993) [21] introduce a class of generalized EM algorithms, which they call the ECM algorithm for expectation-conditional maximization algorithm. The expectation-conditional maximization algorithm (ECM algorithm) takes advantage of the simplicity of complete-data conditional maximization by replacing a complicated M-step of the EM algorithm with several computationally simpler CM-steps. Each of these CM-steps maximizes the conditional expectation of the complete-data log-likelihood function found in the preceding E-step subject to constraints on $\mathbf{\Psi}$, and all constraints are the maximization of the full parameter space of $\mathbf{\Psi}$.

A CM-step may be in the closed form, or it may itself require iteration, but because the CM maximization is over smaller dimensional spaces, often they are simpler, faster, and more stable than the corresponding full maximization called for on the M-step of the EM algorithm, especially when iteration are required.

4.4.2 Formal Definition

To define formally the expectation-conditional maximization algorithm (ECM algorithm), we suppose that the M-step is replaced by $S > 1$ steps. We let $\mathbf{\Psi}^{(k+s/S)}$ denote the value of $\mathbf{\Psi}$ on the sth CM-step of the $(k+1)$th iteration, where $\mathbf{\Psi}^{(k+s/S)}$ is chosen to maximize

$$Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)})$$

subject to the constraint

$$\mathbf{g}_s(\mathbf{\Psi}) = \mathbf{g}_s(\mathbf{\Psi}^{(k+(s-1)/S)}). \tag{4.9}$$

Here $C = \{\mathbf{g}_s(\mathbf{\Psi}), s = 1, \ldots, S\}$ is a set of S preselected (vector) functions. Thus $\mathbf{\Psi}^{(k+s/S)}$ satisfies

$$Q(\mathbf{\Psi}^{(k+s/S)}; \mathbf{\Psi}^{(k)}) \geq Q(\mathbf{\Psi}; \mathbf{\Psi}^{(k)}) \tag{4.10}$$

for all $\mathbf{\Psi} \in \mathbf{\Omega}_s(\mathbf{\Psi}^{(k+(s-1)/S)})$, where

$$\mathbf{\Psi} \in \mathbf{\Omega}_s(\Psi^{(k+(s-1)/S)}) \equiv \{\mathbf{\Psi} \in \mathbf{\Omega} : \mathbf{g}_s(\mathbf{\Psi}) = \mathbf{g}_s(\mathbf{\Psi}^{(k+(s-1)/S)})\}. \tag{4.11}$$

The value of $\mathbf{\Psi}$ on the final CM-step, $\mathbf{\Psi}^{(k+s/S)} = \mathbf{\Psi}^{(k+1)}$, is taken to be input on the $(k+2)$th iteration. Then, we can have that

$$\begin{aligned} Q(\mathbf{\Psi}^{(k+1)}; \mathbf{\Psi}^{(k)}) &\geq Q(\mathbf{\Psi}^{(k+(s-1)/S)}; \mathbf{\Psi}^{(k)}) \\ &\geq Q(\mathbf{\Psi}^{(k+(s-2)/S)}; \mathbf{\Psi}^{(k)}) \\ &\vdots \\ &\geq Q(\mathbf{\Psi}^{(k)}; \mathbf{\Psi}^{(k)}) \end{aligned} \tag{4.12}$$

This shows that the expectation-conditional maximization algorithm (ECM algorithm) is a generalized expectation maximization algorithm and so possesses its desirable convergence properties. The above inequality is a sufficient condition for

$$L(\mathbf{\Psi}^{(k+1)}) \geq L(\mathbf{\Psi}^{(k)})$$

to satisfy.

The sth CM-step then requires the maximization of the Q-function with respect to the sth subvector $\mathbf{\Psi}_s$ with the other $(S-1)$ vectors held fixed at their current values; that is, $\mathbf{g}_s(\mathbf{\Psi})$ is the vector containing all the subvectors of $\mathbf{\Psi}$ except $\mathbf{\Psi}_s$ $(s = 1, \ldots, S)$. [23] In the simulation study, the M-step is not simple, being iterative, and it is replaced by a number of CM-steps which either exist in closed form or require a lower dimensional search.

4.5 Newton-Raphson Method

First, we illustrate the Newton-Raphson method in univariate random variables z and x for solving equations of the form $f(z) = 0$ where $x = f(z)$. We suppose an initial fit for the root we are trying to find and we name it z_0. Consider the $z_0, z_1, \ldots, z_{n+1}$. The equation of the tangent to the graph $x = f(z)$ at the point $(z_0, f(z_0))$ is

$$x - f(z_0) = f'(z_0)(z - z_0). \tag{4.13}$$

The tangent intersects the z-axis when $x = 0$ and $z = z_1$, so

$$-f(z_0) = f'(z_0)(z_1 - z_0). \tag{4.14}$$

Solving the above equation for z_1 gives

$$z_1 = z_0 - \frac{f(z_0)}{f'(z_0)} \tag{4.15}$$

and, so we can have a more general form

$$z_{n+1} = z_n - \frac{f(z_0)}{f'(z_n)} \tag{4.16}$$

The Newton-Raphson method used to solve the likelihood equation, $\mathbf{S}(\mathbf{x}; \mathbf{\Psi}) = \mathbf{0}$, approximates the gradient vector $\mathbf{S}(\mathbf{x}; \mathbf{\Psi})$ of the log-likelihood function $\log L(\mathbf{\Psi})$. We can show

$$\mathbf{S}(\mathbf{x}; \mathbf{\Psi}) \approx \mathbf{S}(\mathbf{x}; \mathbf{\Psi}^{(k)}) - I(\mathbf{\Psi}^{(k)})(\mathbf{\Psi} - \mathbf{\Psi}^{(k)}). \tag{4.17}$$

A new fit $\Psi^{(k+1)}$ is obtained by equating the right-hand side to zero. From (4.17), the iterative equation is obtained as

$$\mathbf{\Psi}^{(k+1)} = \mathbf{\Psi}^{(k)} + I^{-1}(\mathbf{\Psi}^{(k)}; \mathbf{x})\mathbf{S}(\mathbf{x}; \mathbf{\Psi}^{(k)}). \tag{4.18}$$

If we want that the sequence of iterates $\mathbf{\Psi}^{(k)}$ converges to the MLE of $\mathbf{\Psi}$, we should guarantee that the log-likelihood function is concave and unimodal. When the log-likelihood function is not concave, the Newton-Raphson method may not converge given an arbitrary starting value. Under reasonable assumptions on $L(\mathbf{\Psi})$ and a sufficiently accurate starting value, the sequence of iterates $\mathbf{\Psi}^{(k)}$ produced by the Newton-Raphson method can have local quadratic convergence to a solution $\mathbf{\Psi}^*$ of $\mathbf{S}(\mathbf{x}; \mathbf{\Psi}) = \mathbf{0}$. That is, when we give a norm $\| \cdot \|$ on parameter field $\mathbf{\Omega}$, there is a constant h such that if $\mathbf{\Psi}^{(0)}$ is sufficiently close to Ψ^*, then $\| \mathbf{\Psi}^{(k+1)} - \mathbf{\Psi}^* \| \leq h \| \mathbf{\Psi}^{(k)} - \mathbf{\Psi}^* \|$ where $k = 0, 1, 2 \ldots$.

Since quadratic convergence is ultimately very fast, it is regarded as the major strength of the Newton-Raphson method. But there can be potentially serious problems with this method in applications. Firstly, it requires at each iteration, the

computation of the $n \times n$ information matrix $\mathbf{I}(\mathbf{\Psi}^{(k)}; x)$, which is the negative of the Hessian matrix. Furthermore, the basic Newton-Raphson method needs some impractically accurate initial values for $\mathbf{\Psi}$ for the sequence of iterates $\mathbf{\Psi}^{(k)}$ to converge to a solution of $\mathbf{S}(\mathbf{x}; \mathbf{\Psi}) = \mathbf{0}$. It has the trend to head toward saddle points and local minima as often as toward local maxima.

Since the Newton-Raphson method requires $\mathbf{I}(\mathbf{\Psi}^{(k)}; \mathbf{x})$ on each iteration k, it can give an estimate of the covariance matrix of its limiting value Ψ^* assuming that it is the MLE, through the inverse of the observed information matrix $\mathbf{I}^{-1}(\mathbf{\Psi}^{(k)}; \mathbf{x})$.

Chapter 5

Simulation Study

5.1 Generated Skew-normal Data

In this chapter, we conducted a series of simulations to study the differences in parameters' estimates obtained by the skew-normal method and ordinary normal method when true distributions are normal, gamma, chi-square or t in linear mixed models. First, we study empirical properties of the skewness parameter, λ, for different values of the variance components. We also get estimates of parameters (including fixed effects and random effects), their averages, biases, mean squared errors and empirical coverage probabilities.

The data were generated from the following linear mixed model

$$y_{ij} = \alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij} \tag{5.1}$$

where random effects b_i were generated from one of the four distributions under study. We assume that the random errors e_{ij} follow normal distribution, that is $e_{ij} \sim N(0, \sigma_e^2)$.

For our model, we have 5 subjects in each cluster (or group) i, for $i = 1, \ldots, m$, $j = 1, \ldots, 5$. We assume that b_i and e_{ij} are independent. Thus, we can give the

variance and covariance of y_{ij} as we have introduced in Chapter 2, that is

$$\begin{aligned} var(y_{ij}) &= var(\alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij}) \\ &= \sigma_b^2 + \sigma_e^2 \end{aligned}$$

For the covariance of the inter-group,

$$\begin{aligned} cov(y_{ij}, y_{ij'}) &= cov(\alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij}, \alpha + t_{ij'}\beta_1 + w_i\beta_2 + b_i + e_{ij'}) \\ &= \sigma_b^2 \end{aligned}$$

For the covariance of the intra-group,

$$\begin{aligned} cov(y_{ij}, y_{i'j}) &= cov(\alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij}, \alpha + t_{i'j}\beta_1 + w_i'\beta_2 + b_i' + e_{i'j}) \\ &= 0 \end{aligned}$$

Similarly, $cov(y_{ij}, y_{i'j'}) = 0$

For the observed data of fixed effects, (t_{ij}, w_i), we set $t_{ij} = j - 1$, like $t_{i1} = 0$, $t_{i2} = 1$, $t_{i3} = 2$, $t_{i4} = 3$, $t_{i5} = 4$ and set $w_i = 1$ if $i \leq m/2$ and $w_i = 0$ if $i > m/2$, where m is cluster size. We also set intercept $\alpha = 5$, $\beta_1 = 2$, $\beta_2 = 0.5$ for regression parameters.

We can rewrite model (5.1) in the matrix form as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i. \tag{5.2}$$

The variance-covariance matrix of each group is obtained as

$$cov(\mathbf{Y}_i) = \mathbf{V}_b = \begin{pmatrix} \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma_e^2 \end{pmatrix}.$$

The variance-covariance matrix for the whole response vector $\mathbf{Y}$ is obtained as

$$var(\mathbf{Y}) = \begin{pmatrix} \mathbf{V}_b & 0 & \cdots & 0 & 0 \\ 0 & \mathbf{V}_b & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \mathbf{V}_b & 0 \\ 0 & 0 & \cdots & 0 & \mathbf{V}_b \end{pmatrix}.$$

5.2 Likelihood Ratio Test for skewness parameter

In this section, we give a general definition of the likelihood ratio test. Then we derive the likelihood ratio test statistic for testing the significance of skewness parameter λ. Here we test the null hypothesis is $H_0 : \lambda = 0$ against the alternative hypothesis $H_1 : \lambda \neq 0$.

We usually use a proper statistical hypothesis test to see if parameters are significant enough for a particular model. For example, we sometimes use a z-test to assess whether a specific value μ_0 is a plausible scalar for the population mean μ, and we sometimes use a t-test to see whether a specific value μ_0 is a plausible value for the population mean μ. Here, we introduce the likelihood ratio method, which is widely used in hypothesis tests. [19]

Assume $\boldsymbol{\theta}$ to be a vector which consists of all unknown population parameters. Assume that $\mathbf{L}(\boldsymbol{\theta})$ is the likelihood function obtained by evaluating the joint density of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ at their observed values $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. The parameter vector $\boldsymbol{\theta}$ takes its value in the parameter set $\boldsymbol{\Theta}$. For example, in the q-dimensional multivariate normal case, $\boldsymbol{\theta} = [\mu_1, \ldots, \mu_q, \sigma_{11}, \ldots, \sigma_{1q}, \sigma_{22}, \ldots, \sigma_{2q}, \ldots, \sigma_{q-1,q}, \sigma_{qq}]$ and $\boldsymbol{\Theta}$ includes the q-dimensional space, where $-\infty < \mu_1 < \infty, \ldots, -\infty < \mu_q < \infty$ combined with the $q(q+1)/2$-dimensional space of variances and covariances such that $\boldsymbol{\Sigma}$ is positive definite. So, $\boldsymbol{\Theta}$ has dimension $q + q(q+1)/2$. Under the null hypothesis H_0:$\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\boldsymbol{\theta}$ is restricted to lie in a subset $\boldsymbol{\Theta}_0$ of $\boldsymbol{\Theta}$. For the multivariate normal

situation with $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}$ unspecified, $\boldsymbol{\Theta}_0 = \{\mu_1 = \mu_{10}, \mu_2 = \mu_{20}, \ldots, \mu_p = \mu_{p0}; \sigma_{11}, \ldots, \sigma_{1q}, \sigma_{22} \ldots, \sigma_{2q}, \ldots, \sigma_{q-1,q}, \sigma_{qq}\}$ with $\boldsymbol{\Sigma}$ positive definite, so $\boldsymbol{\Theta}_0$ has dimension $q(q+1)/2$. [19]

A likelihood ratio test of H_0: $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ rejects H_0 in favor of H_1: $\boldsymbol{\theta} \subsetneq \boldsymbol{\Theta}_0$ if

$$\boldsymbol{\Lambda} = \frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbf{L}(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{L}(\boldsymbol{\theta})} < c, \tag{5.3}$$

where c is a properly chosen constant value. Intuitively, we reject H_0 if the maximum of the likelihood obtained by allowing $\boldsymbol{\theta}$ to vary over the set $\boldsymbol{\Theta}_0$ is much smaller than the maximum of the likelihood obtained by varying $\boldsymbol{\theta}$ over all values in $\boldsymbol{\Theta}$. When the maximum in the numerator of expression above is much smaller than the maximum in the denominator, $\boldsymbol{\Theta}_0$ does not include proper values for $\boldsymbol{\theta}$.

In each application of the likelihood ratio method, we should obtain the sampling distribution of the likelihood-ratio test statistic $\boldsymbol{\Lambda}$. Then constant value c can be selected to produce a test with a specified significance level α. However, if we can satisfy certain regularity conditions and ensure large cluster size, the sampling distribution of $-2 \ln \boldsymbol{\Lambda}$ is well approximated by a chi-square distribution. This attractive feature is the reason of popularity of likelihood ratio procedures.

Under the null hypothesis, when n is large, we can have that

$$-2 \ln \boldsymbol{\Lambda} = -2 \ln \left(\frac{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbf{L}(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{L}(\boldsymbol{\theta})} \right)$$

is approximately a χ^2_q random variable. Here the degrees of freedom are $q + q(q+1)/2 - q(q+1)/2 = q$. [19]

We now derive the log-likelihood ratio statistic for the skew-normal distribution in order to assess the significance of the skewness parameter λ. Recall that our skew-normal density function is

$$f_{\mathbf{Y}_i}(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}_b) = 2\phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)\Phi_1(\bar{\boldsymbol{\lambda}}_{b_i}^T \boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})) \tag{5.4}$$

with

$$\bar{\lambda}_{b_i} = \frac{\boldsymbol{\Sigma}_i^{-1/2}\mathbf{Z}_i\mathbf{D}^{1/2}\lambda_b}{\sqrt{1+\lambda_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}_i\mathbf{D}^{-1/2}\lambda_b}}$$

$$\boldsymbol{\Lambda}_i = (\mathbf{D}^{-1} + \mathbf{Z}_i^T\boldsymbol{\psi}_i^{-1}\mathbf{Z}_i)^{-1}$$

and

$$\boldsymbol{\Sigma}_i = \boldsymbol{\psi}_i + \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T$$

Since we have

$$\mathbf{Z}_i = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{5\times 1},$$

$$\mathbf{D} = Var(b_i) = \sigma_b^2,$$

$$\boldsymbol{\psi}_i = \sigma_e^2\mathbf{I}_5,$$

we can show that

$$\begin{aligned}\boldsymbol{\Lambda}_i &= (\mathbf{D}^{-1} + \mathbf{Z}_i^T\boldsymbol{\psi}_i^{-1}\mathbf{Z}_i)^{-1} \\ &= (\frac{1}{\sigma_b^2} + \frac{5}{\sigma_e^2})^{-1},\end{aligned}$$

and

$$\boldsymbol{\Sigma}_i = \sigma_e^2\mathbf{I}_5 + \sigma_b^2\mathbf{Z}_i\mathbf{Z}_i^T = \begin{pmatrix} \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_e^2 \end{pmatrix}.$$

We can also simplify $\bar{\lambda}_{b_i}$ as

$$\begin{aligned}\bar{\lambda}_{b_i} &= \frac{\boldsymbol{\Sigma}_i^{-1/2}\mathbf{Z}_i\mathbf{D}^{1/2}\lambda_b}{\sqrt{1+\lambda_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}_i\mathbf{D}^{-1/2}\lambda_b}} \\ &= \frac{\boldsymbol{\Sigma}_i^{-1/2}\mathbf{1}\sigma_b\lambda_b}{\sqrt{1+\lambda_b^2(\frac{1}{\sigma_b^2})(\frac{1}{\sigma_b^2}+\frac{5}{\sigma_e^2})^{-1}}}.\end{aligned}$$

The skew-normal density function can be rewritten as

$$\begin{aligned}f_{\mathbf{Y}_i}(\mathbf{y}_i \mid \boldsymbol{\theta}, \boldsymbol{\lambda}_b) &= 2\phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)\Phi_1(\bar{\boldsymbol{\lambda}}_{b_i}^T\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})) \\ &= 2\phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)\Phi_1\left(\frac{\lambda_b\sigma_b\mathbf{1}^T\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\Sigma}_i^{-1/2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\sqrt{1+\lambda_b^T\mathbf{D}^{-1/2}\boldsymbol{\Lambda}_i\mathbf{D}^{-1/2}\lambda_b}}\right) \\ &= 2\phi_{n_i}(\mathbf{y}_i \mid \mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)\Phi_1\left(\frac{\lambda_b\sigma_b\mathbf{1}^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})}{\sqrt{1+\lambda_b^2(\frac{1}{\sigma_b^2})(\frac{1}{\sigma_b^2}+\frac{5}{\sigma_e^2})^{-1}}}\right).\end{aligned}$$

When we remove the skewness parameter λ, the skew-normal model becomes the normal model. Recall that the joint density function of the linear mixed model under normality is

$$f(\mathbf{y}_i) = \frac{1}{(2\pi)^{n/2} \mid \boldsymbol{\Sigma}_i \mid^{1/2}} exp\left\{-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}. \tag{5.5}$$

Thus, the log-likelihood function is given by

$$l(\boldsymbol{\beta}, \boldsymbol{\varphi}) = -\frac{1}{2}\sum_{i=1}^{m}\log(\mid \boldsymbol{\Sigma}_i \mid) - \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) - \frac{m}{2}\log 2\pi \tag{5.6}$$

Based on this log-likelihood under $H_0 : \lambda = 0$ and the general skew-normal log-likelihood under $H_1 : \lambda \neq 0$, we can conduct a likelihood ratio test for testing H_0 against H_1.

5.3 Empirical levels and powers of likelihood ratio tests for testing the skewness parameter λ

Here we study the empirical properties of the likelihood ratio test for assessing the significance of λ at a given level of significance. The significance level of a test is defined as the probability of rejecting the null hypothesis when it is true. The empirical level of the likelihood ratio test is the proportion of times we reject the null hypothesis H_0 when H_0 is true. In our case, the empirical level that we wish to investigate is the proportion of times we reject the null hypothesis H_0:$\lambda = 0$ when data are generated from the normal distribution, that is, when H_0 is true.

In order to obtain the empirical level of the likelihood ratio test for different cluster sizes and variance components, we performed the following steps.

(i) First, we generated the data from the linear mixed model (5.1) with $b_i \sim N(0, \sigma_b^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$, where the variance components were chosen as $\sigma_b^2 = 8, 4, 2, 1$ and $\sigma_e^2 = 0.5$ respectively.

(ii) Second, we repeated the above to generate a series of 1000 datasets, whose cluster sizes are $m = 20$, $m = 40$, $m = 60$ and $m = 100$. We then performed the likelihood ratio test for testing $H_0 : \lambda = 0$ in each dataset.

(iii) Finally, we calculate the proportion of times that the null hypothesis $H_0 : \lambda = 0$ is rejected, when, in fact, $H_0 : \lambda = 0$ (which means that $b_i \sim N(0, 1)$) is true. We also calculated the empirical powers of the tests under the alternative hypothesis that the random effects b_i has a non-normal chi-square distribution (that is, under $H_A : \lambda \neq 0$) with 3 degrees of freedom.

Table 5.1: Empirical levels and powers for different cluster sizes and variances of random effects and errors

σ_b^2	cluster size	$m = 20$	$m = 40$	$m = 60$	$m = 100$
8	empirical levels	0.039	0.033	0.032	0.025
	powers	0.645	0.952	0.986	1
4	empirical levels	0.04	0.044	0.03	0.022
	powers	0.552	0.797	0.819	0.831
2	empirical levels	0.028	0.033	0.039	0.041
	powers	0.543	0.733	0.745	0.761
1	empirical levels	0.016	0.038	0.025	0.022
	powers	0.522	0.713	0.718	0.724

Table 5.1 presents the empirical levels of the likelihood ratio tests. It is clear from the table that the empirical levels are generally close to the nominal 5% level of significance irrespective of the values of the random effects variance component σ_b^2 and error variance σ_e^2. Also, the empirical powers of the test increases as the number of clusters m increases. The powers also increase when the variance component σ_b^2 increases.

5.4 Estimation in Skew-normal mixed model using EM algorithm

Assuming that only the random effects follow the skew-normal distribution in the linear mixed model:

$$y_{ij} = \alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij}, \quad (i = 1, \ldots, m, j = 1, \ldots, 5) \tag{5.7}$$

where

$$b_i \sim SN(0, \sigma_b^2, \lambda_b), \quad e_{ij} \sim N(0, \sigma_e^2).$$

We also assume that the random errors follow the normal distribution, that is $e_{ij} \sim N(0, \sigma_e^2)$.

For simplicity, we rewrite the above model in matrix form as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i b_i + \mathbf{e}_i \tag{5.8}$$

$$b_i \sim SN(0, \sigma_b^2, \lambda_b), \quad \mathbf{e}_i \sim N_5(\mathbf{0}, \boldsymbol{\psi}_i).$$

We can use the above expression jointly with the additive representation introduced in Section 3.3.1 to obtain

$$b_i = \sigma_b\delta_b \mid X_{0i} \mid + \sigma_b\sqrt{1-\delta_b^2}X_{1i}, \tag{5.9}$$

where $X_{0i} \sim N(0,1)$, $X_{1i} \sim N(0,1)$, with X_{0i} and X_{1i} being independent for $i = 1, \ldots, m$, and $\delta_b = \frac{\lambda_b}{\sqrt{1+\lambda_b^2}}$. Moreover, b_i and $\mathbf{e}_i$ are independent, $i = 1, \ldots, m$. Hence model (5.8) can be rewritten as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \sigma_b\delta_b|X_{0i}|\mathbf{Z}_i + \sigma_b\sqrt{I-\delta_b^2}X_{1i}\mathbf{Z}_i + \mathbf{e}_i, \tag{5.10}$$

where

$$X_{0i} \sim N(0,1), \quad X_{1i} \sim N(0,1), \quad \mathbf{e}_i \sim N_5(\mathbf{0}, \boldsymbol{\psi}_i),$$

with X_{0i}, X_{1i} and $\mathbf{e}_i$ all independent.

For simplicity, we have the following model and assumptions in most cases,

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \bar{\delta}_b t_i \mathbf{Z}_i + \mathbf{r}_i, \tag{5.11}$$

where

$$\bar{\delta}_b = \sigma_b \delta_b, \quad t_i = |X_{0i}|, \quad \mathbf{r}_i = \mathbf{e}_i + \sigma_b \sqrt{1-\delta_b^2} X_{1i} \mathbf{Z}_i,$$

which are such that

$$\mathbf{r}_i \sim N_5(\mathbf{0}, \boldsymbol{\psi}_i + (\sigma_b^2 - \bar{\delta}_b^2)\mathbf{Z}_i\mathbf{Z}_i^T), \quad t_i \sim HN(0,1) \tag{5.12}$$

and are independent, $i = 1, \ldots, m$.

Therefore, (5.7) implies that the conditional model can be written as

$$\begin{aligned} \mathbf{Y}_i | t_i &\sim N_5(\mathbf{X}_i\boldsymbol{\beta} + \bar{\delta}_b t_i \mathbf{Z}_i, \boldsymbol{\psi}_i + (\sigma_b^2 - \bar{\delta}^2)\mathbf{Z}_i\mathbf{Z}_i^T) \\ &\sim N_5(\boldsymbol{\mu}_i + \mathbf{d}_i t_i, \boldsymbol{\Psi}_i) \end{aligned} \tag{5.13}$$

where

$$\begin{aligned} t_i \sim HN_1(0,1), \qquad \boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta}, \quad \mathbf{d}_i = \bar{\delta}_b \mathbf{Z}_i, \quad \boldsymbol{\Psi}_i = \boldsymbol{\Sigma}_i - \mathbf{d}_i\mathbf{d}_i^T, \\ \boldsymbol{\Sigma}_i = \boldsymbol{\psi}_i + \sigma_b^2 \mathbf{Z}_i\mathbf{Z}_i^T, \end{aligned}$$

and d_i and $\boldsymbol{\Psi}_i$ have been shown before. "HN_1" is the half-normal distribution as we introduced earlier. To be more clear, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the marginal mean vector and

covariance matrix, and $\mathbf{\Sigma}_i$ can be presented in the following matrix form,

$$\mathbf{\Psi}_i = \mathbf{\Sigma}_i - \mathbf{d}_i\mathbf{d}_i^T$$
$$= \begin{pmatrix} \sigma_b^2 + \sigma_e^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} \\ \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 + \sigma_e^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} \\ \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 + \sigma_e^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} \\ \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 + \sigma_e^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} \\ \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} & \sigma_b^2 + \sigma_e^2 - \frac{\sigma_b^2\lambda_b^2}{1+\lambda_b^2} \end{pmatrix}$$

In order to get the marginal means of t_i and t_i^2, we first introduce the following two lemmas.

Lemma 5.1 Let $X \sim N(\eta, \tau^2)$. We can use the density function ϕ and distribution function Φ to derive the following formula,

$$E[X|X > a] = \eta + \frac{\phi_1\left(\frac{a-\eta}{\tau}\right)}{1 - \Phi_1\left(\frac{a-\eta}{\tau}\right)}\tau \tag{5.14}$$

$$E[X^2|X > a] = \eta^2 + \tau^2 + \frac{\phi_1\left(\frac{a-\eta}{\tau}\right)}{1 - \Phi_1\left(\frac{a-\eta}{\tau}\right)}(\eta + a)\tau \tag{5.15}$$

for any real constant.

Proof: To give the proof of Lemma 5.1, we will introduce the definition of Truncated Normal Distributions. A random variable X has a doubly truncated normal distribution if its probability density function is [20]

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \left[\frac{1}{\sqrt{2\pi}\sigma}\int_a^b e^{-(t-\mu)^2/2\sigma^2} dt\right]^{-1} \\ &= \sigma^{-1}\phi\left(\frac{x-\mu}{\sigma}\right)\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]^{-1}. \end{aligned} \tag{5.16}$$

The lower and upper truncation points are a, b, respectively; the degrees of truncation are $\Phi((a-\mu)/\sigma)$ (from below) and $1-\Phi((b-\mu)/\sigma)$ (from above). If a is replaced by $-\infty$, or b by ∞, the distribution is singly truncated from above, or below, respectively. It can be shown that when the truncations are large, the distribution bears little resemblance to a normal distribution. The case $a=\mu$, $b=\infty$ produces a half-normal distribution. This is actually the distribution of $\mu+\sigma \mid U \mid$, where U is a unit normal variable. [3]

The moment generating function is [20]

$$\begin{aligned} M(t) &= E[e^{tX} \mid X \in A] \\ &= \frac{\int_a^b e^{tx} f(x)dx}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \\ &= e^{\mu t+\sigma^2 t^2/2} \frac{\Phi((b-\mu)/\sigma - \sigma t) - \Phi((a-\mu)/\sigma - \sigma t)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}. \end{aligned} \tag{5.17}$$

We can get the conditional expected value of X as follows,

$$E[X \mid X \in A] = M'(t) \mid_{t=0} = \mu - \frac{\phi((b-\mu)/\sigma) - \phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}\sigma. \tag{5.18}$$

If $b=\infty$, then [3]

$$E[X \mid X > a] = \mu + \frac{\phi((a-\mu)/\sigma)}{1-\Phi((a-\mu)/\sigma)}\sigma. \tag{5.19}$$

In order to derive the variance, using the moment generating function, [20]

$$\begin{aligned} E[X^2 \mid X \in A] &= M''(t) \mid_{t=0} \\ &= \sigma^2 + \mu^2 + \sigma^2 \frac{\phi'((b-\mu)/\sigma) - \phi'((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi(a-\mu)/\sigma} \\ &- 2\mu\sigma \frac{\phi((b-\mu)/\sigma) - \phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma - \Phi((a-\mu)/\sigma))}. \end{aligned} \tag{5.20}$$

If $a = \mu$, $b = \infty$, then [3]

$$E[X^2|X > a] = \eta^2 + \tau^2 + \frac{\phi_1\left(\frac{a-\eta}{\tau}\right)}{1 - \Phi_1\left(\frac{a-\eta}{\tau}\right)}(\eta + a)\tau.$$

We can also derive the variance as

$$\begin{aligned} Var(X) = \sigma^2 + \sigma^2 \frac{((a-\mu)/\sigma)Z((a-\mu)/\sigma) - ((b-\mu)/\sigma)Z((B-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \\ - \left\{ \frac{Z((a-\mu)/\sigma) - Z((b-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \right\}^2 \sigma^2 \end{aligned} \tag{5.21}$$

Lemma 5.2 Under the conditions in Lemma 5.1,

$$E[T^k \mid \mathbf{y}] = E[X^k \mid X > 0], \tag{5.22}$$

where $X \sim N_1(\eta, \tau^2)$, with η and τ^2 as defined before. Particularly,

$$E[T \mid \mathbf{y}] = \eta + \frac{\phi_1\left(\frac{\eta}{\tau}\right)}{\Phi_1\left(\frac{\eta}{\tau}\right)}\tau \tag{5.23}$$

and

$$E[T^2 \mid \mathbf{y}] = \eta^2 + \tau^2 + \frac{\phi_1\left(\frac{\eta}{\tau}\right)}{\Phi_1\left(\frac{\eta}{\tau}\right)}\tau\eta \tag{5.24}$$

Proof: Note that we can write [3]

$$\begin{aligned} E(T^k|\mathbf{y}) &= \int_{-\infty}^{\infty} t^k f(t|\mathbf{y})dt = \frac{1}{f_{\mathbf{Y}}(\mathbf{y}|\theta,\lambda)} \int_{-\infty}^{\infty} t^k f_{\mathbf{Y},T}(\mathbf{y}, t|\theta, \lambda)dt \\ &= \frac{1}{2\phi_n(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1\left(\frac{\eta}{\tau}\right)} \int_{-\infty}^{\infty} t^k 2\phi_n\left(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \phi_1(t \mid \eta, \tau^2)\mathbb{I}\{t > 0\}dt \\ &= \frac{1}{\Phi_1\left(\frac{\eta}{\tau}\right)} \int_0^{\infty} t^k \phi_1(t|\eta, \tau^2)dt = E(X^k|X > 0), \end{aligned} \tag{5.25}$$

where $X \sim N_1(\eta, \tau^2)$ and $\Phi_1\left(\frac{\eta}{\tau}\right) = P(X > 0)$. They follows Lemma 5.1 with $a = 0$ and the fact that

$$\frac{\eta}{\tau} = \frac{\mathbf{d}^T\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{\mu})}{\sqrt{1 + \mathbf{d}^T\boldsymbol{\Psi}^{-1}\mathbf{d}}}.$$

Proposition 5.1 Suppose that $\mathbf{Y} \mid T = t \sim N_n(\boldsymbol{\mu} + \mathbf{d}t, \boldsymbol{\Psi})$ and $T \sim HN_1(0,1)$ (half-normal distribution). Let $\boldsymbol{\Sigma} = \boldsymbol{\Psi} + \mathbf{d}\mathbf{d}^T$. Then the joint distribution of $(\mathbf{Y}^T, T)^T$ can be written as

$$f_{\mathbf{Y},T}(\mathbf{y}, t|\boldsymbol{\theta}, \boldsymbol{\lambda}) = 2\boldsymbol{\phi}_n(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\, \phi_1(t \mid \eta, \tau^2)\mathbb{I}\{t > 0\}, \tag{5.26}$$

where

$$\eta = \frac{\mathbf{d}^T\boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{\mu})}{1 + \mathbf{d}^T\boldsymbol{\Psi}^{-1}\mathbf{d}},$$

and

$$\tau^2 = \frac{1}{1 + \mathbf{d}^T\boldsymbol{\Psi}^{-1}\mathbf{d}}.$$

Notice that the marginal distribution of Y follows from (5.26) after integrating out t, and is given by

$$f_{\mathbf{Y}}(\mathbf{y} \mid \theta, \lambda) = 2\phi_n(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1\left(\frac{\eta}{\tau}\right). \tag{5.27}$$

Proof: Since the joint density of $\mathbf{Y}$ and T is

$$f_{\mathbf{Y},T}(\mathbf{y}, t|\theta, \lambda) = 2\phi_n(\mathbf{y}|\boldsymbol{\mu} + \mathbf{d}t, \boldsymbol{\psi})\, \phi_1(t)\mathbb{I}\{t > 0\},$$

using Lemma 3.1 or Lemma 3.6, we have

$$\phi_n(\mathbf{y}|\boldsymbol{\mu} + \mathbf{d}t, \boldsymbol{\Psi})\phi_1(t) = \phi_n(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\phi_1(t|\eta, \tau^2).$$

Since the marginal density of Y is

$$f_Y(y \mid \theta, \lambda) = 2\phi_n(y \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi_1\left(\frac{\eta}{\tau}\right),$$

the density of (y,t) can be rewritten as

$$f_{\mathbf{Y},T}(\mathbf{y},t|\theta,\lambda) = 2\phi_n\left(\mathbf{y}|\mu,\mathbf{\Sigma}\right)\phi_1(t \mid \eta,\tau^2)\mathbb{I}\{t>0\},$$

Let $\mathbf{y}^T = (\mathbf{y}_1^T,\ldots,\mathbf{y}_m^T)^T$. We can have following joint distribution of $(\mathbf{Y}^T,T)^T$ according to proposition 5.1,

$$\begin{aligned} f_{\mathbf{Y},T}(\mathbf{y},t|\theta,\lambda) &= 2\phi_n(\mathbf{y}|\boldsymbol{\mu}+\mathbf{d}t,\mathbf{\Psi})\phi_1(t)\mathbb{I}\{t>0\} \\ &= 2\phi_n(\mathbf{y}|\boldsymbol{\mu},\mathbf{\Sigma})\phi_1(t|\eta,\tau^2)\mathbb{I}\{t>0\}, \end{aligned} \tag{5.28}$$

where $|\mathbf{\Psi}+\mathbf{d}\mathbf{d}^T||1+\mathbf{d}^T\mathbf{\Psi}^{-1}\mathbf{d}| = |\mathbf{\Psi}|$.

Proposition 5.2 The complete log-likelihood function associated with $(\mathbf{y},\mathbf{t})$ in the SNLMM (5.8) can be written as

$$\begin{aligned} \ell_c(\theta,\lambda_b) \propto &-\frac{1}{2}\sum_{i=1}^{m}\log \mid \mathbf{\Psi}_i \mid -\frac{1}{2}\sum_{i=1}^{m}\{(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\mathbf{\Sigma}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)\} \\ &-\frac{1}{2}\sum_{i=1}^{m}\frac{(t_i-\eta_i)^2}{\tau_i^2}, \end{aligned} \tag{5.29}$$

where

$$\eta_i = \frac{\mathbf{d}_i^T\mathbf{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{1+\mathbf{d}_i^T\mathbf{\Psi}_i^{-1}\mathbf{d}_i}$$

and

$$\tau_i^2 = \frac{1}{1+\mathbf{d}_i^T\mathbf{\Psi}_i^{-1}\mathbf{d}_i}.$$

Letting $\theta_c = (\theta^T,\lambda_b^T)^T$, $\hat{t}_i = E(T_i \mid \hat{\theta}_c, \mathbf{Y}_i = \mathbf{y}_i)$ and $\hat{t_i^2} = E(T_i^2 \mid \hat{\theta}_c, \mathbf{Y}_i = \mathbf{y}_i)$, we obtain from above Lemma 5.1 that [3]

$$\hat{t}_i = \hat{\eta}_i + \frac{\phi_1\left(\frac{\hat{\eta}_i}{\hat{\tau}_i}\right)}{\Phi_1\left(\frac{\hat{\eta}_i}{\hat{\tau}_i}\right)}\hat{\tau}_i \tag{5.30}$$

and

$$\hat{t_i^2} = \hat{\eta}_i^2 + \hat{\tau}_i^2 + \frac{\phi_1\left(\frac{\hat{\eta}_i}{\hat{\tau}_i}\right)}{\Phi_1\left(\frac{\hat{\eta}_i}{\hat{\tau}_i}\right)}\hat{\tau}_i\hat{\eta}_i. \tag{5.31}$$

Now we will show main steps of differentiating the log-likelihood function above. Before that, we give some expressions and do some preparation work. We can show that

$$\begin{aligned}
\frac{(t_i-\eta_i)^2}{\tau_i^2} &= (t_i-\eta_i)^2 \times \frac{1}{\tau_i^2} \\
&= \left[\frac{\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i} - t_i\right]^2 (1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i) \\
&= \left[\frac{\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{\sqrt{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}} - t_i\sqrt{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}\right]^2
\end{aligned}$$

and its first derivatives are

$$\begin{aligned}
\frac{\partial\left[\frac{(t_i-\eta_i)^2}{\tau_i^2}\right]}{\partial\boldsymbol{\beta}} &= \frac{\partial\left[\frac{\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{\sqrt{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}} - t_i\sqrt{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}\right]^2}{\partial\boldsymbol{\beta}} \\
&= 2\left(\frac{\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{\sqrt{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}} - t_i\sqrt{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}\right) \times \frac{\partial\left[\frac{\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{\sqrt{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}}\right]}{\partial\boldsymbol{\beta}} \\
&= \frac{2\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i\left[t_i(1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i) - \mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)\right]}{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i} \\
&= 2t_i\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i - 2\frac{\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i} \\
&= 2t_i\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i - 2\tau_i^2\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i).
\end{aligned}$$

Then the first derivatives of the whole log-likelihood function give the score function

$$\begin{aligned}
\frac{\partial l_c(\theta,\lambda_b)}{\partial\boldsymbol{\beta}} &= \sum_{i=1}^m \mathbf{X}_i^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i) + \sum_{i=1}^m \tau_i^2\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i) - \sum_{i=1}^m t_i\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i \\
&= \sum_{i=1}^m \mathbf{X}_i^T\left[\boldsymbol{\Sigma}_i^{-1} + \tau_i^2\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\right](\mathbf{y}_i-\boldsymbol{\mu}_i) - \sum_{i=1}^m t_i\mathbf{X}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i.
\end{aligned}$$

Let the score function be 0, and then solve the score equation with respect to $\boldsymbol{\beta}$. This

gives the estimator

$$\begin{aligned}\hat{\boldsymbol{\beta}} =&[\sum_{i=1}^{m}\mathbf{X}_i^T(\hat{\boldsymbol{\Sigma}}_i^{-1}+\hat{\tau}_i^2\hat{\boldsymbol{\Psi}}_i^{-1}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i\hat{\boldsymbol{\Psi}}_i^{-1})\mathbf{X}_i]^{-1}\times\\ &\sum_{i=1}^{m}[\mathbf{X}_i^T(\hat{\boldsymbol{\Sigma}}_i^{-1}+\hat{\tau}_i^2\hat{\boldsymbol{\Psi}}_i^{-1}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i^T\hat{\boldsymbol{\Psi}}_i^{-1})\mathbf{y}_i-\hat{t}_i\mathbf{X}_i^T\hat{\boldsymbol{\Psi}}_i^{-1}\hat{\mathbf{d}}_i].\end{aligned} \tag{5.32}$$

We also have

$$\eta_i=\frac{\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)}{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}$$

and

$$\tau_i^2=\frac{1}{1+\mathbf{d}_i^T\boldsymbol{\Psi}_i^{-1}\mathbf{d}_i}.$$

We then have the following EM algorithm:

E-step: Given $\theta_c=\hat{\theta}_c$, compute $\hat{t}_i$ and $\hat{t}_i^2$ for $i=1,\ldots,m$.

M-step:Update $\hat{\theta}_c$ by maximizing $E[\ell_c(\theta_c)\mid y,\hat{\theta}_c]$ over θ_c, which leads to

$$\begin{aligned}\hat{\boldsymbol{\beta}} =&[\sum_{i=1}^{m}\mathbf{X}_i^T(\hat{\boldsymbol{\Sigma}}_i^{-1}+\hat{\tau}_i^2\hat{\boldsymbol{\Psi}}_i^{-1}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i\hat{\boldsymbol{\Psi}}_i^{-1})\mathbf{X}_i]^{-1}\times\\ &\sum_{i=1}^{m}[\mathbf{X}_i^T(\hat{\boldsymbol{\Sigma}}_i^{-1}+\hat{\tau}_i^2\hat{\boldsymbol{\Psi}}_i^{-1}\hat{\mathbf{d}}_i\hat{\mathbf{d}}_i^T\hat{\boldsymbol{\Psi}}_i^{-1})\mathbf{y}_i-\hat{t}_i\mathbf{X}_i^T\hat{\boldsymbol{\Psi}}_i^{-1}\hat{\mathbf{d}}_i]\end{aligned} \tag{5.33}$$

and

$$\hat{\boldsymbol{\nu}}=\arg\max_{\boldsymbol{\nu}}[\ell_c(\hat{\boldsymbol{\beta}},\boldsymbol{\nu})]\quad\text{with }\boldsymbol{\nu}=(\sigma_b^2,\sigma_e^2,\lambda_b)^T$$

where $\ell_c(\hat{\boldsymbol{\beta}},\boldsymbol{\nu})$ is evaluated at updated $\hat{\boldsymbol{\beta}}$, $t_i=\hat{t}_i$ and $t_i^2=\hat{t}_i^2$,$i=1,\ldots,m$.

5.5 Normal method using Newton-Raphson method

Assuming that b_i and $\boldsymbol{\epsilon}_i$ follow normal distribution, we can have the following Linear Mixed Model,

$$\mathbf{Y}_i=\mathbf{X}_i\boldsymbol{\beta}+\mathbf{Z}_ib_i+\boldsymbol{\epsilon}_i \tag{5.34}$$

$$b_i\sim N(0,\sigma_b^2),\quad \boldsymbol{\epsilon}_i\sim N_5(\mathbf{0},\sigma_e^2\mathbf{I}_5)$$

where ϵ_i are independent, then we can derive the variance-covariance matrix for $\mathbf{Y}_i$. And for simplicity, we rewrite the original model as

$$y_{ij} = \alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + \epsilon_{ij}, \quad (i = 1, \ldots, 200; i = 1, \ldots, 5) \tag{5.35}$$

where b_i's are independent normal variables $N(0, \sigma_b^2)$, ϵ_{ij}'s are independent normal variables $N(0, \sigma_\epsilon^2)$, and σ_b^2 and σ_ϵ^2 are independent of each other. Thus, we can give the variance and covariance of y_{ij}, that is

$$\begin{aligned} var(y_{ij}) &= var(\alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + \epsilon_{ij}) \\ &= \sigma_b^2 + \sigma_\epsilon^2 \end{aligned}$$

For the covariance of the inter-group,

$$\begin{aligned} cov(y_{ij}, y_{ij'}) &= cov(\alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + \epsilon_{ij}, \alpha + t_{ij'}\beta_1 + w_i\beta_2 + b_i + \epsilon_{ij'}) \\ &= \sigma_b^2 \end{aligned}$$

For the covariance of the intra-group,

$$\begin{aligned} cov(y_{ij}, y_{i'j}) &= cov(\alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + \epsilon_{ij}, \alpha + t_{i'j}\beta_1 + w_i'\beta_2 + b_i' + \epsilon_{i'j}) \\ &= 0 \end{aligned}$$

Similarly, $cov(y_{ij}, y_{i'j'}) = 0$

While in the matrix form, $cov(\mathbf{Y}_i)$ and $cov(\mathbf{Y})$ have been shown before.

And the density function is [14]

$$f(\mathbf{y}_i \mid \boldsymbol{\mu}_i, \mathbf{V}_i) = \frac{\exp[-\frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)]}{(2\pi)^{N/2} \mid \mathbf{V}_i \mid^{1/2}}$$

Then we can get the log-likelihood function, [14]

$$l = -\frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \boldsymbol{\mu}_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) - \frac{1}{2}\sum_{i=1}^{m}\log \mid \mathbf{V}_i \mid -\frac{m}{2}\log 2\pi$$

And we have known that $\boldsymbol{\mu} = \boldsymbol{\mu}_{(\boldsymbol{\beta})}$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$, $\mathbf{V} = \mathbf{V}_{\boldsymbol{\sigma}^2}$ and $\boldsymbol{\sigma}^2 = (\sigma_b^2, \sigma_\epsilon^2)$ And we can use the first derivatives to get the score function of $\boldsymbol{\beta}$ and $\mathbf{V}$,

$$\begin{aligned}\frac{\partial l}{\partial \boldsymbol{\beta}} &= -\frac{1}{2}\sum_{i=1}^{m}[2\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)]\left(-\frac{\partial \boldsymbol{\mu}_i{}^T}{\partial \boldsymbol{\beta}}\right) \\ &= \sum_{i=1}^{m}\frac{\partial \boldsymbol{\mu}_i{}^T}{\partial \boldsymbol{\beta}}\mathbf{V}_i^{-1}[\mathbf{y}_i - \boldsymbol{\mu}_i]\end{aligned}$$

And the score function for σ_b^2, [14]

$$\begin{aligned}\frac{\partial l}{\partial \sigma_b^2} &= -\frac{1}{2}\sum_{i=1}^{m}(\mathbf{y} - \boldsymbol{\mu}_i)^T\frac{\partial \mathbf{V}_i^{-1}}{\partial \sigma_b^2}(\mathbf{y}_i - \boldsymbol{\mu}_i) - \frac{1}{2}\sum_{i=1}^{m}\frac{\partial \log \mid \mathbf{V}_i \mid}{\partial \sigma_b^2} \\ &= \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \boldsymbol{\mu}_i)^T\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \sigma_b^2}\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) - \frac{1}{2}\sum_{i=1}^{m}tr\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \sigma_b^2}\right)\end{aligned}$$

Similarly, we can also have the score function for σ_ϵ^2, [14]

$$\begin{aligned}\frac{\partial l}{\partial \sigma_\epsilon^2} &= -\frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \boldsymbol{\mu}_i)^T\frac{\partial \mathbf{V}_i^{-1}}{\partial \sigma_\epsilon^2}(\mathbf{y}_i - \boldsymbol{\mu}_i) - \frac{1}{2}\sum_{i=1}^{m}\frac{\partial \log \mid \mathbf{V}_i \mid}{\partial \sigma_\epsilon^2} \\ &= \frac{1}{2}\sum_{i=1}^{m}(\mathbf{y}_i - \boldsymbol{\mu}_i)^T\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \sigma_\epsilon^2}\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) - \frac{1}{2}\sum_{i=1}^{m}tr\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \sigma_\epsilon^2}\right)\end{aligned}$$

In order to get information matrix, we can derive the second derivatives first,

$$\begin{aligned}\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} &= \frac{\partial}{\partial \boldsymbol{\beta}}\left(\frac{\partial l}{\partial \boldsymbol{\beta}^T}\right) \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}\left[\frac{\partial \boldsymbol{\mu}_i{}^T}{\partial \boldsymbol{\beta}}\mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)\right]^T \\ &= \frac{\partial}{\partial \boldsymbol{\beta}}\left[(\mathbf{y}_i - \boldsymbol{\mu}_i)^T\mathbf{V}_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T}\right] \\ &= -\frac{\partial \boldsymbol{\mu}_i{}^T}{\partial \boldsymbol{\beta}}\mathbf{V}_i^{-1}\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}^T} + (\mathbf{y}_i - \boldsymbol{\mu}_i)^T\mathbf{V}_i^{-1}\frac{\partial^2 \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\end{aligned}$$

Thus, we can have the information matrix of β part, [12]

$$-E\left[\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right] = \frac{\partial\boldsymbol{\mu}_i{}^T}{\partial\boldsymbol{\beta}}\mathbf{V}_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}^T}$$

And we can also get the second derivatives of other parts, [14]

$$\begin{aligned}
\frac{\partial^2 l}{\partial\sigma_b^2\partial\boldsymbol{\beta}^T} &= \frac{\partial}{\partial\sigma_b^2}\left(\frac{\partial l}{\partial\boldsymbol{\beta}^T}\right) \\
&= \frac{\partial}{\partial\sigma_b^2}\left[\frac{\partial\boldsymbol{\mu}_i{}^T}{\partial\boldsymbol{\beta}}\mathbf{V}_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)\right]^T \\
&= \frac{\partial}{\partial\sigma_b^2}\left[(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\mathbf{V}_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}^T}\right] \\
&= (\mathbf{y}_i-\boldsymbol{\mu}_i)^T\frac{\mathbf{V}_i^{-1}}{\partial\sigma_b^2}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}^T} \\
&= (\mathbf{y}_i-\boldsymbol{\mu}_i)^T[-\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\mathbf{V}_i^{-1}]\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}^T}
\end{aligned}$$

And we can find that,

$$-E\left[\frac{\partial^2 l}{\partial\sigma_b^2\partial\boldsymbol{\beta}^T}\right] = 0$$

Similarly,

$$-E\left[\frac{\partial^2 l}{\partial\sigma_\epsilon^2\partial\boldsymbol{\beta}^T}\right] = 0$$

As transpositions,

$$-E\left[\tfrac{\partial^2 l}{\partial\boldsymbol{\beta}^T\partial\sigma_b^2}\right] = 0, \quad -E\left[\tfrac{\partial^2 l}{\partial\boldsymbol{\beta}^T\partial\sigma_\epsilon^2}\right] = 0 \tag{5.36}$$

And

$$\begin{aligned}
\frac{\partial^2 l}{\partial\sigma_b^2\partial\sigma_e^2} = &-\frac{1}{2}\Bigg\{tr(-\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2}+\mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\sigma_b^2\sigma_e^2}) \\
&+(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\Bigg[(-1)\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2}\mathbf{V}_i^{-1}+\mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\sigma_b^2\partial\sigma_e^2}\mathbf{V}_i^{-1} \\
&-\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\mathbf{V}_i^{-1}\Bigg](\mathbf{y}_i-\boldsymbol{\mu}_i)\Bigg\}
\end{aligned}$$

Since for any $\mathbf{A}$, we have that

$$\mathbf{E}[(\mathbf{y}_i - \boldsymbol{\mu}_i)^T \mathbf{A}(\mathbf{y}_i - \boldsymbol{\mu}_i)] = tr\{\mathbf{A}\mathbf{E}[(\mathbf{y}_i - \boldsymbol{\mu}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i)^T]\} = tr(\mathbf{A}\mathbf{V})$$

Therefore,

$$\begin{aligned}
-E\left[\frac{\partial^2 l}{\partial\sigma_b^2\partial\sigma_e^2}\right] =& \frac{1}{2}\Bigg\{tr\left(-\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2} + \mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\sigma_b^2\partial\sigma_e^2}\right) \\
&+ tr\left[\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2} - \mathbf{V}_i^{-1}\frac{\partial^2\mathbf{V}_i}{\partial\sigma_b^2\partial\sigma_e^2} + \mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\right]\Bigg\} \\
=& \frac{1}{2}tr\left(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2}\right)
\end{aligned}$$

According to all above, we can have the information matrix, [14]

$$-\mathbf{E}\begin{pmatrix} \frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial(\boldsymbol{\sigma}^2)^T} \\ (\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial(\boldsymbol{\sigma}^2)^T})^T & \frac{\partial^2 l}{\partial(\boldsymbol{\sigma}^2)\partial(\boldsymbol{\sigma}^2)^T} \end{pmatrix} = \begin{pmatrix} \frac{\partial\boldsymbol{\mu}_i^T}{\partial\boldsymbol{\beta}}\mathbf{V}_i^{-1}\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\left\{tr(\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_e^2}\mathbf{V}_i^{-1}\frac{\partial\mathbf{V}_i}{\partial\sigma_b^2})\right\} \end{pmatrix}$$

5.6 Bias, Mean Squared Errors and Empirical Coverage Probability

Here we study the empirical biases, mean squared errors (MSEs) and coverage probabilities of the estimators for model parameters under skew-normal random effects ($i.e, b_i \sim SN(0, \sigma_b^2, \lambda_b)$) and normal random effects ($i.e, b_i \sim N(0, \sigma_b^2)$).

5.6.1 Bias

The bias of an estimator $\hat{\theta}$ of a parameter θ is calculated as the difference between the expected value of $\hat{\theta}$ and the parameter θ, given by

$$\begin{aligned} Bias(\hat{\theta}) &= E(\hat{\theta}) - \theta \\ &\cong \left\{ \frac{1}{S} \sum_{s=1}^{S} \hat{\theta}^{(s)} \right\} - \theta \end{aligned} \tag{5.37}$$

where $\hat{\theta}^{(s)}$ is the estimator of θ obtained at the s^{th} simulation. [2]

We should also note that sometimes positive biases can offset negative biases, so we usually pay more attention to mean squared errors.

5.6.2 Mean Squared Errors

The mean squared error (MSE) of an estimator $\hat{\theta}$ of a parameter θ can be obtained from the set of simulations as

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &\cong \frac{1}{S} \sum_{s=1}^{S} \left(\hat{\theta}^{(s)} - \theta \right)^2 \end{aligned} \tag{5.38}$$

where $\hat{\theta}^{(s)}$ is the estimator of θ obtained at the s^{th} simulation. [2]

5.6.3 Empirical Coverage Probability

Coverage Probability refers to the probability that a procedure for constructing random regions will produce an interval containing, or covering the true value. It is a method of the interval generating process and is independent of the particular sample to which such a process is applied. We can think of the indicator as the probability that the interval constructed by such a procedure will contain the parameters of

interest. That is,

$$P_{EC} = \frac{\text{Number of times the intervals contain true } \theta}{\text{Number of confidence intervals}} \tag{5.39}$$

5.7 Simulation Study Under Different Random Effects

5.7.1 Simulation Study Under Normal Distribution

The normal distribution has a mean parameter μ and a variance parameter σ^2. We generate random effects, b_i, from the normal distribution, which has $E[b_i] = 0$ and $Var[b_i] = 3$. That is,

$$b_i \sim \text{Normal}(\mu = 0, \sigma^2 = 3)$$

Here we try to put α, the intercept, and b_i, the random effects, together. So we have

$$\begin{aligned} E[\alpha + b_i] &= \alpha + E[b_i] = 5, \\ Var[\alpha + b_i] &= Var[b_i] = 3. \end{aligned} \tag{5.40}$$

We generate data from the linear mixed model

$$Y_{ij} = \alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij}, \quad (i = 1, \ldots, m, j = 1, \ldots, 5) \tag{5.41}$$

with

$$b_i \sim N(0, 3), \quad e_{ij} \sim N(0, 0.5).$$

In this case, we study the biases, MSEs and empirical coverage probabilities of the estimators of model parameters under skew-normal and normal mixed models, when the true distribution is the normal distribution, where $b_i \sim N(0, 3)$ and $e_{ij} \sim N(0, 0.5)$.

For the observed data of fixed effects, (t_{ij}, w_i), we set $t_{ij} = j - 1$, so that $t_{i1} = 0$, $t_{i2} = 1$, $t_{i3} = 2$, $t_{i4} = 3$, $t_{i5} = 4$ and set $w_i = 1$ if $i \leq m/2$ and $w_i = 0$ if $i > m/2$, where m is cluster size. We also set intercept $\alpha = 5$, $\beta_1 = 2$, $\beta_2 = 0.5$ for regression parameters, so that $(\alpha, \beta_1, \beta_2) = (5, 2, 0.5)$.

Tables 5.2-5.4 present means, biases, mean squared errors and empirical coverage probabilities of the estimators of model parameters under skew-normal and normal mixed models, when the true distribution is the normal distribution, where $b_i \sim N(0,3)$ and $e_{ij} \sim N(0,0.5)$.

Table 5.2: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is Normal(0,3). Number of clusters m=40.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	1.988	-0.012	0.3275	0.955
$\beta_2(0.5)$	0.5019	0.0019	0.0013	0.951
$\sigma_e^2(0.5)$	0.5295	0.0295	0.006	0.933
$E[\alpha+b](5)$	5.0082	0.0082	0.1773	–
$Var[\alpha+b](3)$	2.8184	-0.1816	0.4624	–
Under Normal fit				
$\beta_1(2)$	1.9906	-0.0094	0.3272	0.953
$\beta_2(0.5)$	0.5019	0.0019	0.0013	0.951
$\sigma_e^2(0.5)$	0.4961	-0.0039	0.0033	0.96
$E[\alpha+b](5)$	5.0069	0.0069	0.1776	–
$Var[\alpha+b](3)$	2.8347	-0.1653	0.4676	–

Table 5.3: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is Normal(0,3). Number of clusters m=60.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	1.9818	-0.0182	0.1916	0.949
$\beta_2(0.5)$	0.4992	-0.0008	0.0009	0.954
$\sigma_e^2(0.5)$	0.5301	0.0301	0.0043	0.932
$E[\alpha+b](5)$	5.0041	0.0041	0.104	–
$Var[\alpha+b](3)$	2.857	-0.143	0.3068	–
Under Normal fit				
$\beta_1(2)$	1.9815	-0.0185	0.1909	0.951
$\beta_2(0.5)$	0.4992	-0.0008	0.0009	0.954
$\sigma_e^2(0.5)$	0.4987	-0.0013	0.0022	0.949
$E[\alpha+b](5)$	5.0042	0.0042	0.1039	–
$Var[\alpha+b](3)$	2.877	-0.123	0.3089	–

Table 5.4: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is Normal(0,3). Number of clusters m=100.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	2.0058	0.0058	0.1191	0.955
$\beta_2(0.5)$	0.4999	-0.0001	0.0005	0.955
$\sigma_e^2(0.5)$	0.5269	0.0269	0.0024	0.908
$E[\alpha + b](5)$	5.0035	0.0035	0.0648	–
$Var[\alpha + b](3)$	2.9473	-0.0527	0.1843	–
Under Normal fit				
$\beta_1(2)$	2.0058	0.0058	0.119	0.955
$\beta_2(0.5)$	0.4999	-0.0001	0.0005	0.955
$\sigma_e^2(0.5)$	0.4968	-0.0032	0.0012	0.953
$E[\alpha + b](5)$	5.0035	0.0035	0.0648	–
$Var[\alpha + b](3)$	2.9704	-0.0296	0.1868	–

It is apparent from Tables 5.2-5.4 that the skew-normal and normal mixed models generally provide unbiased estimates of both the regression parameters and variance components. The biases of the estimators of the regression parameters generally decrease as the number of clusters m increases. Here we should note that there is no big bias for all estimates, no matter under the skew-normal fit or under the normal fit.

For the MSEs of the estimators of model parameters with different cluster sizes, the empirical study shows that the MSEs decrease as the number of clusters m increases. The empirical coverage probabilities are also similar under two fits.

It is clear from the tables that the skew-normal model provides estimates that are almost as efficient as those obtained from the correctly specified normal model. The skew-normal fit is "robust" in that sense.

5.7.2 Simulation Study Under Gamma Distribution

The gamma distribution is one of the most common continuous probability distributions, which belong to two-parameter family distributions. The gamma distribution has a shape parameter k and a scale parameter θ. We can also use the rate parameter which is the inverse of scale parameter, $\beta = 1/\theta$. We can derive the mean and variance of gamma distribution and denote them by a shape parameter k and a scale parameter θ. Here we just present them as $E(x) = k\theta$ and $Var(x) = k\theta^2$. However, the biggest problem for us to conduct a simulation study is that we should make the mean of within-group subjects to be zero, since this is one of the basic assumptions of linear mixed models. Since the shape parameter k and the scale parameter θ need to be positive, we can not set b_i at mean zero. We generate $\alpha + b_i$ from gamma distribution, which has $E[\alpha + b_i] = 5$ and $Var[\alpha + b_i] = 3$. And the intercept, α is constant value. That is,

$$b_i^* \sim \text{Gamma}(\theta = 1, k = 3),$$

which means that we can have the mean and variance as

$$E(b_i^*) = 3, \quad var(b_i^*) = 3.$$

But what we put into our model is b_i, not b_i^*, where we set $b_i = b_i^* - 3$. Here we try to put α, the intercept, and b_i, the random effects, together to make up for $\alpha + b_i$. Thus we can get that

$$\begin{aligned} E[\alpha + b_i] &= \alpha + E[b_i] = 5, \\ Var[\alpha + b_i] &= Var[b_i] = 3. \end{aligned} \tag{5.42}$$

We fit the linear mixed model

$$Y_{ij} = \alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij}, \tag{5.43}$$

assuming b_i follows a skew-normal distribution. We also study the estimators under the normality assumption for b_i.

For the observed data of fixed effects, (t_{ij}, w_i), we set $t_{ij} = j - 1$, so that $t_{i1} = 0$, $t_{i2} = 1$, $t_{i3} = 2$, $t_{i4} = 3$, $t_{i5} = 4$ and set $w_i = 1$ if $i \leq m/2$ and $w_i = 0$ if $i > m/2$, where m is cluster size. We also set intercept $\alpha = 5$, $\beta_1 = 2$, $\beta_2 = 0.5$ for regression parameters, so that $(\alpha, \beta_1, \beta_2) = (5, 2, 0.5)$.

Tables 5.5-5.7 present means, biases, mean squared errors (MSEs) and empirical coverage probabilities of the estimators of model parameters under skew-normal and normal mixed model, when the true distribution is the gamma distribution, where $b_i \sim gamma(3, 1)$ and $e_{ij} \sim N(0, 0.5)$.

Table 5.5: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is gamma(3,1). Number of clusters m=40.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	1.9827	-0.0173	0.2837	0.955
$\beta_2(0.5)$	0.4989	-0.0011	0.0012	0.96
$\sigma_e^2(0.5)$	0.5442	0.0442	0.0088	0.912
$E[\alpha + b](5)$	5.0289	0.0289	0.1536	–
$Var[\alpha + b](3)$	2.8498	-0.1502	0.8576	–
Under Normal fit				
$\beta_1(2)$	1.9844	-0.0156	0.306	0.957
$\beta_2(0.5)$	0.4989	-0.0011	0.0012	0.96
$\sigma_e^2(0.5)$	0.4958	-0.0042	0.003	0.965
$E[\alpha + b](5)$	5.0281	0.0281	0.159	–
$Var[\alpha + b](3)$	2.8799	-0.1201	0.8599	–

It is apparent from Tables 5.5-5.7 that the skew-normal and normal mixed models generally provide unbiased estimates of both the regression parameters and variance component. The biases of the estimators of the regression parameters generally decrease as the number of clusters m increases. Here we should note that there is no big bias for all estimates, no matter under the skew-normal fit or under the normal fit.

For the MSEs of the estimators of model parameters with different cluster sizes, the empirical study shows that the MSEs decrease as the number of clusters m increases. And empirical coverage probabilities are also similar under two fits except variance of errors, σ_e^2.

Table 5.6: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is gamma(3,1). Number of clusters m=60.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	1.9821	-0.0179	0.1767	0.961
$\beta_2(0.5)$	0.5008	0.0008	0.0008	0.953
$\sigma_e^2(0.5)$	0.5482	0.0482	0.0082	0.885
$E[\alpha+b](5)$	4.9904	-0.0096	0.0936	–
$Var[\alpha+b](3)$	2.8213	-0.1787	0.6067	–
Under Normal fit				
$\beta_1(2)$	1.9812	-0.0188	0.1915	0.959
$\beta_2(0.5)$	0.5008	0.0008	0.0008	0.953
$\sigma_e^2(0.5)$	0.499	-0.001	0.002	0.945
$E[\alpha+b](5)$	4.9909	-0.0091	0.0961	–
$Var[\alpha+b](3)$	2.8584	-0.1416	0.6072	–

Here the skew-normal model generally provides more efficient estimates of the regression parameters, as compared to the ordinary normal model. For example, when estimating β_1 for $m = 60$, Table 5.6 shows that the skew-normal model provides an MSE of 0.1767, whereas the normal model provides a larger MSE of 0.1915.

Table 5.7: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is gamma(3,1). Number of clusters m=100.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	2.0118	0.0118	0.1082	0.945
$\beta_2(0.5)$	0.4999	-0.0001	0.0005	0.954
$\sigma_e^2(0.5)$	0.5413	0.0413	0.0061	0.888
$E[\alpha + b](5)$	4.9845	-0.0155	0.0604	–
$Var[\alpha + b](3)$	2.8828	-0.1172	0.3869	–
Under Normal fit				
$\beta_1(2)$	2.0137	0.0137	0.1152	0.949
$\beta_2(0.5)$	0.4999	-0.0001	0.0005	0.954
$\sigma_e^2(0.5)$	0.4979	-0.0021	0.0012	0.954
$E[\alpha + b](5)$	4.9835	-0.0165	0.0614	–
$Var[\alpha + b](3)$	2.919	-0.081	0.385	–

5.7.3 Simulation Under Chi-square Distribution

Here we generated the random effects b_i from

$$b_i = b_i^* - 3, \tag{5.44}$$

where b_i^* is chi-square with 3 degrees of freedom. In this case,

$$\begin{aligned} E[\alpha + b_i] &= \alpha + E[b_i] = 5 \\ Var[\alpha + b_i] &= Var[b_i] = 6 \end{aligned} \tag{5.45}$$

We fit the linear mixed model

$$Y_{ij} = \alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij}, \tag{5.46}$$

assuming that b_i follows a skew-normal distribution. We also study the estimators under the normality assumption for b_i.

For the observed data of fixed effects, (t_{ij}, w_i), we set $t_{ij} = j - 1$, like $t_{i1} = 0$, $t_{i2} = 1$, $t_{i3} = 2$, $t_{i4} = 3$, $t_{i5} = 4$ and set $w_i = 1$ if $i \leq m/2$ and $w_i = 0$ if $i > m/2$, where m is cluster size. We also set intercept $\alpha = 5$, $\beta_1 = 2$, $\beta_2 = 0.5$ for regression parameters, so that $(\alpha, \beta_1, \beta_2) = (5, 2, 0.5)$.

Tables 5.8-5.10 present means, biases, mean squared errors and empirical coverage probabilities of the estimators of model parameters under skew-normal and normal mixed model, when the true distribution is the chi-square distribution, where $b_i \sim chisq(3)$ and $e_{ij} \sim N(0, 0.5)$.

It is apparent from Tables 5.8-5.10 that the skew-normal and normal mixed models generally provide unbiased estimates of both the regression parameters and variance components. The biases of the estimators of the regression parameters generally increase as the number of clusters m increases. We should note that there is no big

Table 5.8: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is chi(3). Number of clusters m=40.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	2.0164	0.0164	0.4153	0.95
$\beta_2(0.5)$	0.4984	-0.0016	0.0013	0.947
$\sigma_e^2(0.5)$	0.6864	0.1864	0.0432	0.986
$E[\alpha + b](5)$	5.0029	0.0029	0.2439	–
$Var[\alpha + b](6)$	5.7444	-0.2556	4.5317	–
Under Normal fit				
$\beta_1(2)$	2.005	0.005	0.5765	0.954
$\beta_2(0.5)$	0.4984	-0.0016	0.0013	0.947
$\sigma_e^2(0.5)$	0.4966	-0.0034	0.0031	0.959
$E[\alpha + b](5)$	5.0086	0.0086	0.288	–
$Var[\alpha + b](6)$	5.5993	-0.4007	4.6454	–

bias for all estimates, no matter under the skew-normal fit or under the normal fit.

For the MSEs of the estimators of model parameters with different cluster sizes, the empirical study shows that the MSEs decrease as the number of clusters m increases. The empirical coverage probabilities are also similar under two fits.

It should be pointed out that here the skew-normal model generally provides more efficient estimates of the regression parameters, as compared to the ordinary normal model. For example, when estimating β_1 for $m = 60$, Table 5.9 shows that the skew-normal model provides an MSE of 0.2842, whereas the normal model provides a much larger MSE of 0.4155.

Table 5.9: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is chi(3). Number of clusters m=60.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	2.0027	0.0027	0.2842	0.956
$\beta_2(0.5)$	0.5004	0.0004	0.0009	0.956
$\sigma_e^2(0.5)$	0.7012	0.2012	0.0459	0.913
$E[\alpha + b](5)$	4.9916	-0.0084	0.171	–
$Var[\alpha + b](6)$	5.7176	-0.2824	3.4039	–
Under Normal fit				
$\beta_1(2)$	1.9982	-0.0018	0.4155	0.948
$\beta_2(0.5)$	0.5004	0.0004	0.0009	0.956
$\sigma_e^2(0.5)$	0.5011	0.0011	0.0021	0.96
$E[\alpha + b](5)$	4.9939	-0.0061	0.2075	–
$Var[\alpha + b](6)$	5.5833	-0.4167	3.4934	–

Table 5.10: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is chi(3). Number of clusters m=100.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	1.9926	-0.0074	0.159	0.946
$\beta_2(0.5)$	0.4989	-0.0011	0.0005	0.953
$\sigma_e^2(0.5)$	0.4992	-0.0008	0.0012	0.957
$E[\alpha + b](5)$	5.0135	0.0135	0.1057	–
$Var[\alpha + b](6)$	5.9032	-0.0968	2.097	–
Under Normal fit				
$\beta_1(2)$	1.9956	-0.0044	0.2501	0.95
$\beta_2(0.5)$	0.4989	-0.0011	0.0005	0.953
$\sigma_e^2(0.5)$	0.7098	0.2098	0.0471	0.971
$E[\alpha + b](5)$	5.012	0.012	0.1302	–
$Var[\alpha + b](6)$	5.7119	-0.2881	2.1669	–

5.7.4 Simulation Under t Distribution

Student's t distribution with k degrees of freedom for a given constant μ can be regarded as the distribution of t with $t = \frac{X+\mu}{Z/k}$, where X is a standard normal with $N(0,1)$; Z has a chi-squared distribution with k degrees of freedom; X and Z are independent. We can also derive the mean and variance denoted by k and μ. That is, $E(x) = \mu$ and $Var(x) = k/(k-2)$. As the consequence, we generated $\alpha + b_i$ from t distribution, which has $E[\alpha + b_i] = 5$ and $Var[\alpha + b_i] = 3$. That is,

$$b_i \sim \mathrm{t}(\mu = 0, df = 3),$$

which means that

$$E(b_i) = 0, \quad var(b_i) = 3.$$

Here we try to put α, the intercept, and b_i together, so that

$$\begin{aligned} E[\alpha + b_i] &= \alpha + E[b_i] = 5, \\ Var[\alpha + b_i] &= Var[b_i] = 3. \end{aligned} \tag{5.47}$$

We fit the linear mixed model to be

$$Y_{ij} = \alpha + t_{ij}\beta_1 + w_i\beta_2 + b_i + e_{ij} \tag{5.48}$$

$$b_i \sim SN_q(0, \sigma_b^2, \lambda_b), \quad \epsilon_{ij} \sim N(0, 0.5)$$

assuming that b_i follows a skew-normal distribution. We also study the estimators under the normality assumption for b_i.

For the observed data of fixed effects, (t_{ij}, w_i), we set $t_{ij} = j - 1$, so that $t_{i1} = 0$, $t_{i2} = 1$, $t_{i3} = 2$, $t_{i4} = 3$, $t_{i5} = 4$ and set $w_i = 1$ if $i \leq m/2$ and $w_i = 0$ if $i > m/2$, where m is cluster size. We also set intercept $\alpha = 5$, $\beta_1 = 2$, $\beta_2 = 0.5$ for regression parameters, so that $(\alpha, \beta_1, \beta_2) = (5, 2, 0.5)$.

Tables 5.11-5.13 present means, biases, mean squared errors and empirical coverage probabilities of the estimators of model parameters under skew-normal and normal mixed model, when the true distribution is the t distribution, where $b_i \sim N(0, 3)$ and $e_{ij} \sim N(0, 0.5)$.

Table 5.11: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is t(3). Number of clusters m=40.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	1.9541	-0.0459	0.2527	0.945
$\beta_2(0.5)$	0.4995	-0.0005	0.0012	0.952
$\sigma_e^2(0.5)$	0.5213	0.0213	0.0058	0.993
$E[\alpha + b](5)$	5.0249	0.0249	0.1273	–
$Var[\alpha + b](6)$	2.6058	-0.3942	4.5148	–
Under Normal fit				
$\beta_1(2)$	1.9558	-0.0442	0.2615	0.947
$\beta_2(0.5)$	0.4995	-0.0005	0.0012	0.952
$\sigma_e^2(0.5)$	0.495	-0.005	0.0031	0.961
$E[\alpha + b](5)$	5.024	0.024	0.1292	–
$Var[\alpha + b](6)$	2.6148	-0.3852	4.5495	–

It is apparent from Tables 5.11-5.13 that the skew-normal and normal mixed models generally provide unbiased estimates of both the regression parameters and variance components. The biases of the estimators of the regression parameters generally decrease and variance of errors σ_e^2 as the number of clusters m increases. And here we should note that there is no big bias for all estimates, no matter under the skew-normal fit or under the normal fit.

For the MSEs of the estimators of model parameters with different cluster sizes, the empirical study shows that the MSEs decrease as the number of clusters m increases. And empirical coverage probabilities are also similar under two fits.

It should be pointed out that here the skew-normal model generally provides more efficient estimates of the regression parameters, as compared to the ordinary normal model. For example, when estimating β_1 for $m = 60$, Table 5.12 shows that the

Table 5.12: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is t(3). Number of clusters m=60.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	1.9747	-0.0253	0.1859	0.955
$\beta_2(0.5)$	0.5002	0.0002	0.0008	0.952
$\sigma_e^2(0.5)$	0.524	0.024	0.0043	0.946
$E[\alpha + b](5)$	5.0155	0.0155	0.0943	–
$Var[\alpha + b](6)$	2.6598	-0.3402	2.7112	–
Under Normal fit				
$\beta_1(2)$	1.9754	-0.0246	0.1899	0.956
$\beta_2(0.5)$	0.5002	0.0002	0.0008	0.952
$\sigma_e^2(0.5)$	0.497	-0.003	0.0022	0.959
$E[\alpha + b](5)$	5.0151	0.0151	0.095	–
$Var[\alpha + b](6)$	2.6754	-0.3246	2.7662	–

skew-normal model provides an MSE of 0.1859, whereas the normal model provides a little larger MSE of 0.1899.

Table 5.13: Empirical biases, MSEs, and coverage probabilities of the ML estimators under the assumption of skew-normal and normal distributions for the random effects, when the true distribution is t(3). Number of clusters m=100.

Parameter	Mean	Bias	MSE	EC
Under Skew-normal fit				
$\beta_1(2)$	2.0188	0.0188	0.1099	0.957
$\beta_2(0.5)$	0.4992	-0.0008	0.0005	0.948
$\sigma_e^2(0.5)$	0.525	0.025	0.003	0.941
$E[\alpha + b](5)$	4.9901	-0.0099	0.0606	–
$Var[\alpha + b](6)$	2.6639	-0.3361	1.668	–
Under Normal fit				
$\beta_1(2)$	2.0174	0.0174	0.1118	0.96
$\beta_2(0.5)$	0.4992	-0.0008	0.0005	0.948
$\sigma_e^2(0.5)$	0.4987	-0.0013	0.0013	0.958
$E[\alpha + b](5)$	4.9908	-0.0092	0.0611	–
$Var[\alpha + b](6)$	2.6836	-0.3164	1.7129	–

Chapter 6

An Application

6.1 Framingham Heart Study

Here we present an analysis of longitudinal data on cholesterol levels collected from the famed Framingham Heart Study. This experiment was based on data collected from different cholesterol levels over time, age at baseline and gender for $m = 200$ randomly selected individuals. The details are given in Zhang and Davidian (2001) [9]. Here we consider the same linear mixed model as used by these authors. We use the following response and covariates in the model.

y_{ij} is cholesterol level divided by 100 at the ith time for subject j

t_{ij} is time measured in years from baseline, which is (time-5)/10

age_i represents baseline age; sex_i is gender indicator (0=female, 1=male). For the given data, we fit three models with different assumptions as follows,

Model 1

$$y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_i + \beta_3 t_{ij} + b_{0j} + b_{1j} t_{ij} + \epsilon_{ij}$$

$$\mathbf{b}_j \sim SN_2(\mathbf{0}, \mathbf{D}, \boldsymbol{\lambda}_b), \quad \epsilon_{ij} \sim N_1(0, \sigma_\epsilon^2)$$

where $\boldsymbol{\lambda}_b = (\lambda_{b1}, \lambda_{b2})^T$ and $\mathbf{b}_j = (b_{0j}, b_{1j})^T$.

Model 1 has an independent normal distribution for the errors and a multivariate skew-normal distribution for random effects.

Model 2

$$y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_i + \beta_3 t_{ij} + b_{0j} + b_{1j} t_{ij} + \epsilon_{ij}$$

$$\mathbf{b}_i \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \epsilon_{ij} \sim SN_1(0, \sigma^2, \lambda)$$

Model 2 has an independent skew-normal distribution for random errors with a common variance σ_ϵ^2 and a multivariate normal distribution for the random effects.

Model 3

$$y_{ij} = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{age}_i + \beta_3 t_{ij} + b_{0j} + b_{1j} t_{ij} + \epsilon_{ij}$$

$$\mathbf{b}_i \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \epsilon_{ij} \sim N_1(0, \sigma^2)$$

Model 3 is a purely Gaussian model. We can also have our matrix $\mathbf{x}_{ij} = (1, \text{age}_i, \text{sex}_i, t_{ij})^T$, $\mathbf{b}_j = (b_{0j}, b_{1j})^T$ and $\mathbf{Z}_i = (1, t_{ij})^T$. Zhang and Davidian (2000) [9] studied these data and found that the data followed some asymmetric property.

6.2 Analysis of the Framingham Heart Data

For Models 1-3, we assumed $\boldsymbol{\psi}_i = \sigma^2 \mathbf{I}_{n_i}, i = 1, ..., 200$ which means conditional independence for the outcome variable. The following table gives us the estimates from fitting the three models presented earlier. We use the EM-type algorithm which was introduced in Chapter 4 to get the Hessian matrix that can help us to obtain estimated asymptotic SE.

The AIC, BIC and HQ criteria show us that Model 1 under skew-normal assumption presents the best fit. Asymmetry is not detected in Model 2, which is assumed to have asymmetrically distributed random errors, since all parameter estimates are close to the ones in Model 3. Estimates of the individual-level covariate coefficients β_1, β_2 and β_3 are somewhat different in three models.

Table 6.1: Results of fitting models 1, 2 and 3 to the cholesterol data collected as part of the famed Framingham Heart Study

	Model 1		Model 2		Model 3	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
β_0	1.3755	0.1418	1.5955	0.1568	1.5968	0.1543
β_1	-0.0591	0.0534	-0.063	0.0554	-0.063	0.0568
β_2	0.0136	0.0034	0.0184	0.0035	0.0184	0.0037
β_3	0.2281	0.0511	0.2817	0.024	0.2817	0.0242
σ_e^2	0.0434	0.0025	0.0434	0.0024	0.0434	0.0024
$d11$	0.5601	0.0418	0.3716	0.0199	0.3715	0.0201
$d12$	0.0701	0.0317	0.0563	0.0173	0.0563	0.0179
$d22$	0.1924	0.0311	0.1868	0.0308	0.1868	0.0329
λ_{b1}	2.9947	0.7789	–	–	–	–
λ_{b2}	0.0001	0.4814	–	–	–	–
λ_e	–	–	0.014	0.989	–	–
log-likelihood	666.1717		659.7560		659.7560	
AIC	0.6285		0.6233		0.6243	
BIC	0.6057		0.6020		0.6053	
HQ	0.6195		0.6152		0.6171	

Chapter 7

Conclusion

The purpose of the thesis was to study the empirical properties of the likelihood ratio test for assessing the significance of skewness parameter λ at a given level of significance. We also study the biases, MSEs and empirical coverage probabilities of the estimators of model parameters under skew-normal and normal mixed model, when the true distribution is the normal, gamma, chi-square and t distribution.

The empirical levels of the likelihood ratio tests are generally close to the nominal 5% level of significance irrespective of the values of the random effects variance component σ_b^2 and error variance σ_e^2. Also, the empirical powers of the test increases as the number of clusters m increases. The powers also increase when the variance component σ_b^2 increases.

It is apparent that the skew-normal and normal mixed models generally provide unbiased estimates of both the regression parameters and variance component. The biases of the estimators of the regression parameters generally decrease as the number of clusters m increases. Here we should note that there is no big bias for all estimates, no matter under the skew-normal fit or under the normal fit. When comparing estimates of two fits, we can see that biases are very close to each other.

For the MSEs of the estimators of model parameters with different cluster sizes, the empirical study shows that the MSEs decrease as the number of clusters m increases.

And empirical coverage probabilities are also similar under two fits except variance of errors, σ_e^2. That is, when the "true" random effects distributions are skewed, the assumption of a skew-normal distribution for the random effects provides more robust estimators of the model parameters in terms of smaller biases and mean squared errors, as compared to the assumption of a normal distribution for the random effects.

Chapter 8

Appendix

8.1 Code for empirical levels

```
library(nlme)
library(mvtnorm)

linmm.dat <- function(N=100, beta=c(5, 2, 0.5),
sigmasq.b=1, sigmasq.e=0.5)
{
  patient.id <- c(1:N)
  treat <- rep(c(0, 1), c(N/2, N/2))
  b <- rnorm(N, mean=0, sd=sqrt(sigmasq.b))
  dat <- NULL
  for(i in 1:N)
  {
    t0 <- seq(0, 4, 1)
    n <- length(t0)
    x <- rep(treat[i], n)
    mu <- beta[1] + beta[2] * x + beta[3] * t0
```

```
    id <- rep(patient.id[i], n)
    e <- rnorm(n, mean=0, sd=sqrt(sigmasq.e))
    b0 <- rep(b[i], n)
    y <- mu + b0 + e
    dat <- rbind(dat, cbind(id, t0, x, y, b0))
  }
  dat
}


linmm.dat1 <- function(N=100, beta=c(5, 2, 0.5),
sigmasq.b=1, sigmasq.e=0.5)
{
  patient.id <- c(1:N)
  treat <- rep(c(0, 1), c(N/2, N/2))
  b <- rchisq(N, df=3)
  dat <- NULL
  for(i in 1:N)
  {
    t0 <- seq(0, 4, 1)
    n <- length(t0)
    x <- rep(treat[i], n)
    mu <- beta[1] + beta[2] * x + beta[3] * t0
    id <- rep(patient.id[i], n)
    e <- rnorm(n, mean=0, sd=sqrt(sigmasq.e))
    b0 <- rep(b[i], n)
    y <- mu + b0 + e
    dat <- rbind(dat, cbind(id, t0, x, y, b0))
```

```
  }
  dat
}

linmm.like.sn <- function(dat=data0, beta=c(5, 2, 0.5),
sigmasq.b=1, sigmasq.e=0.5, delta=0)
{
  dat <- data.frame(dat)
  initial <- c(beta, sigmasq.b, sigmasq.e, delta)
  unique.id <- unique(dat$id)
  N <- length(unique.id)
  obj.fn <- function(theta)
  {
    beta <- theta[1:3]
    sigmasq.b <- theta[4]
    sigmasq.e <- theta[5]
    delta <- theta[6]
    log.like <- 0
    for (i in 1:N)
    {
      yi <- dat$y[dat$id==unique.id[i]]
      n <- length(yi)
      ti <- dat$t0[dat$id==unique.id[i]]
      xi  <- dat$x[dat$id==unique.id[i]]
      xxi <- cbind(1, xi, ti)
      mui <- c(xxi %*% beta)
      ZDZ <- sigmasq.b * matrix(1, n, n)
      Sigma <- sigmasq.e * diag(rep(1, n))
```

```
        V <- ZDZ + Sigma
        fi <- dmvnorm(yi, mean=mui, sigma=V)
        Lambda <- 1/(1/sigmasq.b + n/sigmasq.e)
        Z <- rep(1, n)
        lambda.b <- delta/(1-delta^2)
        den <- (1/(lambda.b^2) + 1/(1 + n * sigmasq.b/sigmasq.e))
        Q <- sqrt(sigmasq.b) * t(Z) %*% solve(V) %*% (yi-mui)/sqrt(den)
        Phi <- pnorm(c(Q), mean=0, sd=1)
        log.likei <- log(2) + log(fi) + log(Phi)
        log.like <- log.like + log.likei
      }
      # print(-log.like)
      -log.like
    }
    like.fit <- optim(par = initial, fn=obj.fn, hessian=TRUE,
    method = c("L-BFGS-B"),
    lower=c(-Inf, -Inf, -Inf, 0.1, 0.1, -1),
    upper=c(Inf, Inf, Inf, 25, 2, 1))
    cat(like.fit$message, "\n")
    estimate <- round(like.fit$par, digits=4)
    std.err <- round(sqrt(diag(solve(like.fit$hessian[1:5, 1:5]))),
    digits=4)
    objective <- like.fit$value
    list(estimate=estimate, std.err=std.err, objective=objective)
}

emp.level <- function(sim=2, N=40, beta=c(5, 2, 0.5),
sigmasq.b=1, sigmasq.e=0.5)
```

```
{
  q0 <- qchisq(0.95, 1)
  test1 <- NULL
  for(s in 1:sim)
  {
    cat(s)
    data0 <- linmm.dat(N=N, beta=beta, sigmasq.b=sigmasq.b,
    sigmasq.e=sigmasq.e)
    ml.fit <- lme(y~x+t0, data=data.frame(data0), random=~1|id,
    method="ML")
    sn.fit <- linmm.like.sn(dat=data0, beta=beta, sigmasq.b=sigmasq.b,
    sigmasq.e=sigmasq.e, delta=0.01)
    L0 <- ml.fit$logLik
    L1 <- -sn.fit$objective
    T0 <- 2 * (L1-L0)
    test0 <- ifelse(T0 > q0, 1, 0)
    test1 <- rbind(test1, c(T0, test0))
  }
  cat("\n")
  test1
}
```

```
emp.power <- function(sim=2, N=40, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)
{
  q0 <- qchisq(0.95, 1)
```

```
  test1 <- NULL
  for(s in 1:sim)
  {
    cat(s)
    data0 <- linmm.dat1(N=N, beta=beta, sigmasq.b=sigmasq.b,
    sigmasq.e=sigmasq.e)
    ml.fit <- lme(y~x+t0, data=data.frame(data0), random=~1|id,
    method="ML")
    sn.fit <- linmm.like.sn(dat=data0, beta=beta, sigmasq.b=sigmasq.b,
    sigmasq.e=sigmasq.e, delta=0.01)
    L0 <- ml.fit$logLik
    L1 <- -sn.fit$objective
    T0 <- 2 * (L1-L0)
    test0 <- ifelse(T0 > q0, 1, 0)
    test1 <- rbind(test1, c(T0, test0))
  }
  cat("\n")
  test1
}

# N=100

emp1 <- emp.level(sim=1000, N=100, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)
pow1 <- emp.power(sim=1000, N=100, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)

# N=60
```

```
emp1 <- emp.level(sim=1000, N=60, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)
pow1 <- emp.power(sim=1000, N=60, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)

# N=40

emp1 <- emp.level(sim=1000, N=40, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)
pow1 <- emp.power(sim=1000, N=40, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)

# N=20

emp1 <- emp.level(sim=1000, N=20, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)
pow1 <- emp.power(sim=1000, N=20, beta=c(5, 2, 0.5), sigmasq.b=1,
sigmasq.e=0.5)
```

8.2 Code for estimates for different distributions

```
library(nlme)
library(mvtnorm)

# Code for generating data

linmm.dat1 <- function(N=100, beta=c(5, 2, 0.5), sigmasq.b=1,
```

```
sigmasq.e=0.5)
{
  patient.id <- c(1:N)
  treat <- rep(c(0, 1), c(N/2, N/2))
  b <- rchisq(N, df=3) # for generating from chi-square(3)
  dat <- NULL
  for(i in 1:N)
  {
    t0 <- seq(0, 4, 1)
    n <- length(t0)
    x <- rep(treat[i], n)
    mu <- beta[1] + beta[2] * x + beta[3] * t0
    id <- rep(patient.id[i], n)
    e <- rnorm(n, mean=0, sd=sqrt(sigmasq.e))
    b0 <- rep(b[i], n)
    y <- mu + b0 + e
    dat <- rbind(dat, cbind(id, t0, x, y, b0))
  }
  dat
}


linmm.like.sn <- function(dat=data0, beta=c(5, 2, 0.5), sigmasq.b=1,

sigmasq.e=0.5, delta=0.1)
{
  dat <- data.frame(dat)
  initial <- c(beta, sigmasq.b, sigmasq.e, delta)
```

```
unique.id <- unique(dat$id)
N <- length(unique.id)
obj.fn <- function(theta)
{
  beta <- theta[1:3]
  sigmasq.b <- theta[4]
  sigmasq.e <- theta[5]
  delta <- theta[6]
  log.like <- 0
  for (i in 1:N)
  {
    yi <- dat$y[dat$id==unique.id[i]]
    n <- length(yi)
    ti <- dat$t0[dat$id==unique.id[i]]
    xi  <- dat$x[dat$id==unique.id[i]]
    xxi <- cbind(1, xi, ti)
    mui <- c(xxi %*% beta)
    ZDZ <- sigmasq.b * matrix(1, n, n)
    Sigma <- sigmasq.e * diag(rep(1, n))
    V <- ZDZ + Sigma
    fi <- dmvnorm(yi, mean=mui, sigma=V)
    Z <- rep(1, n)
    lambda.b <- log(delta/(1-delta))
    den <- (1/(lambda.b^2) + 1/(1 + n * sigmasq.b/sigmasq.e))
    Q <- sqrt(sigmasq.b) * t(Z) %*% solve(V) %*% (yi-mui)/sqrt(den)
    Phi <- pnorm(c(Q), mean=0, sd=1)
    log.likei <- log(2) + log(fi) + log(Phi)
    log.like <- log.like + log.likei
```

```
    }
    -log.like
  }

  like.fit <- optim(par = initial, fn=obj.fn, hessian=TRUE,
  method = c("L-BFGS-B"),
  lower=c(-Inf, -Inf, -Inf, 0.1, 0.1, 0.01),
  upper=c(Inf, Inf, Inf, 25, 2, 0.99))
  cat(like.fit$message, "\n")
  estimate <- round(like.fit$par, digits=4)
  std.err <- round(sqrt(diag(solve(like.fit$hessian[1:5, 1:5]))),
  digits=4)
  objective <- like.fit$value
  list(estimate=estimate, std.err=std.err, objective=objective)
}

EM<-function(N=N,x,y,beta0=sn.fit$estimate[1:3],D=sn.fit$estimate[4],
resie=sn.fit$estimate[5],delta=sn.fit$estimate[6])
{
  beta0<-lm(y~x-1)$coef
  theta0<-c(beta0,D,resie,delta,4.5)
  scor<-1
  beta<-theta0[1:3]
  J <- matrix(c(1),5,5)
  be<-y-x%*%beta
  mu<-x%*%beta
  part1<-matrix(data=0,nrow=5,ncol=5)
  part2<-matrix(data=0,nrow=3,ncol=3)
```

```
part3<-matrix(data=0,nrow=3,ncol=1)
infdelee1<-0
tj1<-0
tjsq1<-0
err<-1
d<-delta*c(rep(1,5))*sqrt(D)
sigma<-diag(resie,5)+D*c(rep(1,5))%*%t(c(rep(1,5)))
fy<-sigma-d%*%t(d)
taosq<-1/(1+t(d)%*%solve(fy)%*%d)
tao<-sqrt(taosq)
for(i in 1:200){
  eta<-(t(d)%*%solve(fy)%*%(y[(5*i-4):(5*i)]-mu[(5*i-4):(5*i)]))*taosq
  tj<-eta+dnorm(eta/tao)/pnorm(eta/tao)*tao
  tjsq<-eta^2+taosq+dnorm(eta/tao)/pnorm(eta/tao)*tao*eta
  tj1<- c(tj1, tj)
  tjsq1<- c(tjsq1, tjsq)
}
while(err > 0.0001)
{
  beta<-theta0[1:3]
  be<-y-x%*%beta
  mu<-x%*%beta
  part1<-matrix(data=0,nrow=5,ncol=5)
  part2<-matrix(data=0,nrow=3,ncol=3)
  part3<-matrix(data=0,nrow=3,ncol=1)
  infdelee1<-0
  fy1<-0
  scor1<-0
```

```
scor2<-0
d<-delta*c(rep(1,5))*sqrt(D)
sigma<-diag(resie,5)+D*c(rep(1,5))%*%t(c(rep(1,5)))
fy<-sigma-d%*%t(d)
fy1<- c(fy1, diag(fy))
taosq<-1/(1+t(d)%*%solve(fy)%*%d)
tao<-sqrt(taosq)
for(i in 1:N){
  eta<-(t(d)%*%solve(fy)%*%(y[(5*i-4):(5*i)]

  -mu[(5*i-4):(5*i)]))*taosq
  part11<-solve(sigma)+as.numeric(taosq)*solve(fy)%*%d

  %*%t(d)%*%solve(fy)
  part22<-t(x[(5*i-4):(5*i),])%*%part11%*%x[(5*i-4):(5*i),]
  part2<-part2+part22
  part33<-t(x[(5*i-4):(5*i),])%*%part11%*%y[(5*i-4):(5*i)]

  -as.numeric(tj1[i+1])*t(x[(5*i-4):(5*i),])%*%solve(fy)%*%d
  part3<-part3+part33
  scor11<-t(x[(5*i-4):(5*i),])%*%part11%*%(y[(5*i-4):(5*i)]

  -mu[(5*i-4):(5*i)])-as.numeric(tj1[i+1])*t(x[(5*i-4):(5*i),])

  %*%solve(fy)%*%d
  scor1<-scor1+scor11
  scorcon<-(as.numeric(eta)-as.numeric(tj1[i+1]))*t(d)

  %*%solve(fy)%*%solve(fy)%*%(2*mu[(5*i-4):(5*i)]

  -2*y[(5*i-4):(5*i)]+as.numeric(tj1[i+1])*d+as.numeric(eta)*d)
  scor22<-0.5*t(y[(5*i-4):(5*i)]-mu[(5*i-4):(5*i)])
```

```
  %*%solve(sigma)%*%solve(sigma)%*%(y[(5*i-4):(5*i)]

  -mu[(5*i-4):(5*i)])-0.5*scorcon-0.5*sum(diag(solve(fy)))
  scor2<-scor2+scor22
  difetadelta<-t(d)%*%solve(fy)%*%solve(fy)%*%(y[(5*i-4):(5*i)]

  -mu[(5*i-4):(5*i)])/(1+t(d)%*%solve(fy)%*%d)+t(d)%*%solve(fy)

  %*%solve(fy)%*%d%*%t(d)%*%solve(fy)%*%(y[(5*i-4):(5*i)]

  -mu[(5*i-4):(5*i)])/((1+t(d)%*%solve(fy)%*%d)^2)
  infdelcon<-t(d)%*%solve(fy)%*%solve(fy)%*%(2*mu[(5*i-4):(5*i)]

  -2*y[(5*i-4):(5*i)]+2*as.numeric(eta)*d)*difetadelta+t(d)

  %*%solve(fy)%*%solve(fy)%*%solve(fy)%*%((-2*as.numeric(eta)^2

  +2*as.numeric(tj1[i+1])^2)*d+4*(as.numeric(eta)

  -as.numeric(tj1[i+1]))*(y[(5*i-4):(5*i)]-mu[(5*i-4):(5*i)]))
  infdelee11<-sum(diag(solve(sigma)%*%solve(sigma)))

  +0.5*infdelcon-0.5*sum(diag(solve(fy)%*%solve(fy)))
  infdelee1<-infdelee1+infdelee11
}
scor<-c(scor1,scor2)
beta<-solve(part2)%*%part3
ve <- resie + solve(infdelee1)*scor2
vb<-var(y-x%*%beta)-ve
malb<-mean(y-x%*%beta)+beta[1]
theta1<-c(beta,vb,ve,mean(delta),malb)
err<-sqrt(sum(theta1-theta0)^2)
theta0<-theta1
```

```
  }
  return(c(theta1))
}

NR <- function(x,y)
{
  beta0<-c(1,1,1)
  theta0<-c(3,0.15)
  I <- diag(5)
  J <- matrix(c(1),5,5)
  beta <- NULL
  theta <- NULL
  score<-1
  parameter<-c(beta0,theta0)
  while(sum(score^2) > 0.001)
  {
    v0 <- parameter[5]*I+parameter[4]*J
    v0inv <- solve(v0)
    scorebeta <- 0
    scorethetau <- 0
    scorethetae<-0
    fisherbeta <- matrix(data=0,nrow=3,ncol=3)
    infor<-matrix(data=5,nrow=2,ncol=2)
    fisherthetauu <- 200/2 * sum(diag(v0inv %*% J %*% v0inv %*% J))
    fisherthetaue <- 200/2 * sum(diag(v0inv %*% J %*% v0inv))
    fisherthetaeu <- 200/2 * sum(diag(v0inv %*% v0inv %*% J))
    fisherthetaee <- 200/2 * sum(diag(v0inv %*% v0inv))
    infor<-matrix(c(fisherthetauu,fisherthetaue,fisherthetaeu,
```

```
    fisherthetaee),nrow=2,ncol=2,byrow=TRUE)
    for(i in 1:200)
    {
      ysj <- y[(5*i-4):(5*i)]
      xsj <- x[(5*i-4):(5*i),]
      r <- ysj - xsj %*% beta0
      scorebeta0 <- t(xsj) %*% v0inv %*% r
      scorebeta <- scorebeta + c(scorebeta0)
      fisherbeta0 <- t(xsj) %*% v0inv %*% xsj
      fisherbeta <- fisherbeta + fisherbeta0
      scorethetau0 <- -1/2 * sum(diag(v0inv%*%J)) + 1/2 * t(r)
     %*% v0inv %*% J %*% v0inv %*% r
      scorethetau <- scorethetau + scorethetau0
      scorethetae0 <- -1/2 * sum(diag(v0inv)) + 1/2 * t(r) %*%
      v0inv %*% v0inv %*% r
      scorethetae <- scorethetae + scorethetae0
    }
    beta1 <- beta0 + solve (fisherbeta) %*% scorebeta
    scoretheta<-c(scorethetau,scorethetae)
    theta1 <- theta0 + solve (infor) %*%scoretheta
    score<-c(scorebeta,scoretheta)
    parameter<-c(beta1,theta1)
    beta0<-parameter[1:3]
    theta0<-parameter[4:5]
  }
  return(c(parameter))
}
```

```
N=60
beta=c(5, 2, 0.5)
sigmasq.b=2.6
sigmasq.e=0.5
EM1<-NULL
EM2<-NULL
EM3<-NULL
EM4<-NULL
EM5<-NULL
EM6<-NULL
EM7<-NULL
NR1<-NULL
NR2<-NULL
NR3<-NULL
NR4<-NULL
NR5<-NULL

for(s in 1:1000)
{
  data0 <- linmm.dat1(N=N, beta=beta, sigmasq.b=sigmasq.b,
  sigmasq.e=sigmasq.e)
  sn.fit <- linmm.like.sn(dat=data0, beta=beta, sigmasq.b=sigmasq.b,
  sigmasq.e=sigmasq.e, delta=0.1)
  para<-EM(N=N,x=cbind(1,data0[,3],data0[,2]),y=data0[,4],
  beta0=sn.fit$estimate[1:3],D=1,resie=sn.fit$estimate[5],
  delta=sn.fit$estimate[6])
```

```
  EM10<-para[1]
  EM20<-para[2]
  EM30<-para[3]
  EM40<-para[4]
  EM50<-para[5]
  EM60<-para[6]
  EM70<-para[7]
  EM1<- c(EM1, EM10)
  EM2<- c(EM2, EM20)
  EM3<- c(EM3, EM30)
  EM4<- c(EM4, EM40)
  EM5<- c(EM5, EM50)
  EM6<- c(EM6, EM60)
  EM7<- c(EM7, EM70)
  paramt<-NR(x=cbind(1,data0[,3],data0[,2]),y=data0[,4])
  NR10<-paramt[1]
  NR20<-paramt[2]
  NR30<-paramt[3]
  NR40<-paramt[4]
  NR50<-paramt[5]
  NR1<- c(NR1, NR10)
  NR2<- c(NR2, NR20)
  NR3<- c(NR3, NR30)
  NR4<- c(NR4, NR40)
  NR5<- c(NR5, NR50)
}

print(cbind(EM1,EM2,EM3,EM4,EM5,EM7,NR1,NR2,NR3,NR4,NR5))
```

Bibliography

[1] Azzalini A. and Dalla Valle A. The multivariate skew-normal distribution. *Biometrika*, 83:715–726, 1996.

[2] Alia Alkhathami. *Score tests for testing homogeneity of recurrent event times using frailty models*. PhD thesis, Carleton University, 2015.

[3] Bolfarine H. Arellano-Valle R. B. and Lachos V. H. Skew-normal linear mixed models. *Journal of Data Science*, 3:415–438, 2005.

[4] Bolfarine H. Arellano-Valle R. B. and Lachos V. H. Bayesian inference for skew-normal linear mixed models. *Journal of Applied Statistics*, 43:663682, 2007.

[5] Barry C. Arnold and Robert J. Beaver. Skewed multivariate models related to hidden truncation and/or selective reporting. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 11:7–54, 2002.

[6] A. Azzalini. A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12:171178, 1985.

[7] Adelchi Azzalini. *The skew-normal and related families*. Cambridge University Press, 2014.

[8] Marcia D. Branco and Dipak K. Dey. A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, 79:99–113, 2001.

[9] Zhang D. and Davidian M. Liner mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57:795–802, 2001.

[10] G. S. Datta and P. Lahiri. A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statist. Sinica*, 10:613627, 2000.

[11] Laird N. M. Dempster A. P. and Rubin D. B. Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. R. Stat. Soc.*, 39:138, 1977.

[12] Liang K. Y. Diggle P. J. and Zeger S. L. *Analysis of Longitudinal Data*. Oxford Univ. Press, 1987.

[13] B. E. Ellison. Two theorems for inferences about the normal distribution with applications in acceptance sampling. *J. Amer. Statist. Assoc.*, 59:8995, 1964.

[14] Charles E.Mcculloch and Shayle R. Searle. *Generalized, Linear, and Mixed Models*. John Wiley and Sons, Inc, 2001.

[15] M. Ghosh and J.N.K. Rao. Small area estimation: An appraisal (with discussion). *Statist. Sci.*, 9:5593, 1994.

[16] H. O. Hartley and J. N. K. Rao. Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54:93108, 1967.

[17] Norbert Henze. A probabilistic representation of the 'skew-normal' distribution. *Scandinavian Journal of Statistics*, 13:271–275, 1986.

[18] Jiming Jiang. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer Science and Business Media, LLC, 2007.

[19] Richard A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education, Inc, 2007.

[20] Kotz.S. Johnson.N. and Balakrishnan.N. *Continuous Univariate Distribution Vol. 1.* John Wiley and Sons, Inc, 1995.

[21] Meng X. L. and Rubin D. B. Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika*, 80:267278, 1993.

[22] E.L. Lehmann and G Casella. *Theory of Point Estimation.* Springer Verlag, 2007.

[23] Geoffrey J. McLachlan and Thriyamkam Krishnan. *The EM Algorithm and Extensions (Second Edition).* John Wiley and Sons, Inc, 2008.

[24] G.L. McLachlan and K.E. Basford. *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, 1988.

[25] J. J. Miller. Multivariate symmetry and asymmetry. *Encyclopedia of Statistical Sciences*, 8:53385345, 2006.