A Skew-Normal Copula-Driven Generalized Linear Mixed Model for Longitudinal Data

Mohamad Elmasri

Department of Mathematics and Statistics

Concordia University

Montreal, QC

Presented in Partial Fulfillment of the Requirement for the Degree of

Master of Science in Mathematics

April 2012

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mohamad Elmasri**

Entitled: **A Skew-Normal Copula-Driven Generalized Linear Mixed Model For Longitudinal Data**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science in Mathematics

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____________________ Examiner

Dr. C. Hyndman

____________________ Examiner

Dr. Y. P. Chaubey

____________________ Thesis Supervisor

Dr. A. Sen

Approved by ____________________

Chair of Department or Graduate Program Director

____

Dean of Faculty

Date ____________________

Abstract

A Skew-Normal Copula-Driven Generalized Linear Mixed Model for Longitudinal Data

Mohamad Elmasri

Using the advancements of Arellano-Valle et al. [2005], which characterize the likelihood function of a linear mixed model (LMM) under a skew-normal distribution for the random effects, this thesis attempt to construct a copula-driven generalized linear mixed model (GLMM). Assuming a multivariate distribution from the exponential family for the response variable and a skew-normal copula, we drive a complete characterization of the general likelihood function. For estimation, we apply a Monte Carlo expectation maximization (MC-EM) algorithm. Some special cases are discussed, in particular, the exponential and gamma distributions. Simulations with multiple link functions are shown alongside a real data example from the Framingham Heart Study.

Acknowledgements

I would like to thank Professor Kalyan Das, Department of Statistics, University of Calcutta for the proposition of the model, and my supervisor, professor Arusharka Sen, for his patience explaining the model, and his general guidance and knowledge that have aided me greatly throughout my undergraduate and graduate studies. I am grateful to Professor Yogendra Chaubey for his guidance and support during my graduate studies and for reviewing this work. I am also grateful to Professor Lea Popovic for all the discussions that guided me generally and for the countless reference letters she wrote for me.

I am very grateful to all the friends I have, for the unconditional support, for all the good times we have spent together, for the cheers and laughs we shared, and for baring with me during the hard times. Here I must mention a few; Santiago, for the endless hours we spent in the lab during my studies; Catherine and Maryam, for their warmth and smiles, and for putting up with me during the hard times. Without exception, I am fortunate to have your friendship.

Special thanks to my uncle Ghassan and his wife Racha, for their continuous support and generosity.

To my parents, Maha and Ahmed, and to my two lovely sisters

Fatima and Maryam.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The linear mixed model

The key component driving the development of linear mixed models is the ability of such models to handle data with non-independent observations; a data structure where predictor/response variables are measured at more than one level. Such structure is common with repeated observations, notably longitudinal data in medical studies where patient characteristics are measured over varying times. Because of the imposed dependence between observations from the same source, the presumed independence of errors in linear models is in turn violated and the Ordinary Least Square method fails to capture the characteristics of coefficients.

The first improvement on the linear model to accommodate hierarchical data was proposed by Fisher [1918], discussing the correlation between relatives through Mendelian inheritance. Fisher proposed the addition of the random effects term to the linear model, which in turn relaxed the homoscedastic condition on error. Later, Robinson [1991] provided best linear unbiased estimates (BLUE) of the fixed effects and best linear prediction (BLUP).

To characterize the linear mixed model, define the different measurement levels as units, and let $\mathbf{Y_i}$ be an $(n_i \times 1)$ of observed response variable for sample unit i, $i = 1, ..., m$. Then $\mathbf{Y_i}$ is defined as

$$\mathbf{Y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}b_i + \epsilon_\mathbf{i} \ , \ i = 1, \dots, m \tag{1.1.1}$$

where $\mathbf{X_i}$ of dimension $(n_i \times p)$ is the design matrix corresponding to the fixed effects β of dimension $(p \times 1)$, $\mathbf{Z_i}$ of dimension $(n_i \times q)$ is the design matrix that incorporates the hierarchical variables as random effects, $\mathbf{b_i}$ is the random effects regression coefficients of dimension $(q \times 1)$ and $\epsilon_\mathbf{i}$ of dimension $(n_i \times 1)$ is the vector of random errors. Note that the terms in $\mathbf{Z_i}$ represent a non-time variant or categorical variables that constitute the hierarchical structure of the data.

Inferences from this model become slightly more tedious by the addition of the random coefficient b_i to the known error terms $\epsilon_\mathbf{i}$. A general approach to manage such complexity is by assuming independence between b_i and $\epsilon_\mathbf{i}$ as follow

$$b_i \overset{iid}{\sim} N_q(0, D), \quad \epsilon_\mathbf{i} \overset{ind}{\sim} N_{ni}(0, \psi_i) \tag{1.1.2}$$

where $D = D(\alpha)$, $\psi_i = \psi_i(\gamma)$, for all $i = 1, \dots, m$ are associated dispersion matrices depending on reduced parameters $\alpha \ \ and \ \ \gamma$ with possible variability among -and within- individuals. Putting aside the independence assumption between random effects and residual, the extra restrictiveness associated with distribution function characteristics and structure of both b_i and ϵ_i is deemed to be unnecessary in many literature reviews. Although Butler and Louis [1992] have recently shown that the normality assumption has little effect on the fixed effects estimates, this assumptions' effect on the random effects estimates has not been investigated until Verbeke and Lesaffre [1996]. They demonstrated the use of a mixture of normals in estimating the random effects coefficients by iterative means using the EM algorithm. Although their method has widely expanded the boundaries of model estimation the drawbacks are more apparent when observations depart from the normality assumption.

On the other hand, Zhang and Davidian [2001] adopted another approach using the semi-parametric form in estimating random effects by extending Gallant and Nychka [1987] development of maximum likelihood semi-parametric estimation procedures. Another iterative technique was demonstrated by Tao et al. [1999], where they extended the work of Magder and Zeger [1996] by a predictive technique, they also compared non-parametric maximum likelihood (NPMLE) and smoothed non-parametric maximum likelihood(SNPMLE) to Newton and Zhang [1999] predictive recursion algorithm (PR). Finally, Arellano-Valle et al. [2005], expanded the boundaries of the normally distributed random effects coefficients and error to a skew-normal distribution, where the skew-normal characteristics enhance the presumed distributions to include any slight departures caused by skewness. Arellano-Valle et al. [2005]'s work was facilitated by the work of Azzalini [1985], which constructed the distribution of a univariate skew-normal via an additive mixture of normal and half normal distributions. Arellano-Valle et al. [2005] expands on such concepts to deal with multivariate situations, where they have explicitly characterized the likelihood function and used a constrained expectation maximization algorithm (CEM) to systematically produce coefficient estimates.

This thesis takes a new approach to modeling hierarchical multivariate distributions. In particular, given response variables $Y_{ij}, \quad i = 1, \ldots n,\ j = 1, \ldots, n_i$, we assume that Y_i follows n_i-variate distribution with a predefined mean and variance-covariance matrix. We will model such distribution by using a n_i-variate skew-normal copula $SN_{n_i}(.)$ and integrating the random effects in the mean structure of the copula. Furthermore, we chose the variance-covariance matrix Σ_i to be of an autoregressive structure in order to include the time-variant parameters. Formally,

$$Y_i | b_i \sim MV(\eta(X_i\beta + b_i), \Sigma_i(\phi_i, t_i)) \tag{1.1.3}$$

where b_i is the unit specific unobserved random effect, ϕ_i is the dispersion autoregressive time-variant parameter and $\eta(.)$ is a link function.

The thesis is organized as follow; Chapter 2 discusses the formation of the univariate and multivariate skew-normal distribution. Chapter 3 introduces the model and constructs the likelihood using a skew-normal copula within a GLM framework. Chapter 4 discusses the use of numerical Monte Carlo EM-algorithm to estimate parameters of the established model. Chapter 5 presents numerical simulations and a real data example using the proposed model.

1.2 The skew-normal distribution

This thesis uses the following notations; $\phi_n(.|\mu, \Sigma)$ and $\Phi_n(.|\mu, \Sigma)$ to represent a n-variate normal probability density and distribution functions respectively, with location vector μ and scale $(n \times n)$ variance-covariance matrix Σ. Hence, for $\mu = 0$ and $\Sigma = I_n$ the previous notations would simplify to an n-variate standard normal probability density and distribution functions $\phi_n(.)$ and $\Phi_n(.)$ respectively. In addition, let $SN_n(.|\mu, \Sigma, \lambda)$ and $sn_n(.|\mu, \Sigma, \lambda)$ be a n-variate skew-normal distribution and density functions respectively, with skewness vector λ. Similarly, a n-variate standard skew-normal distribution and density functions are represented by $SN_n(.|\lambda)$ and $sn_n(.|\lambda)$ respectively. Finally, let $HN(.)$ be the half normal distribution function.

The following two sections summarize previous work and advancements in modeling a univariate and multivariate skew-normal distributions.

1.2.1 Univariate skew-normal distribution

Following Azzalini [1985], a random variable X has a skew-normal distribution with skewness parameter λ if density function is represented as

$$sn_1(x|\lambda) = 2\phi_1(x)\Phi_1(\lambda x) \quad (-\infty < x < \infty) \tag{1.2.1}$$

similarly by

$$sn_1(x|\mu, \sigma^2, \lambda) = 2\phi_1(x|\mu, \sigma^2)\Phi_1(\lambda \frac{x - \mu}{\sigma}) \tag{1.2.2}$$

$$(-\infty < x < \infty), \quad \mu, \sigma \in \Re, \quad \mu < \infty, 0 < \sigma < \infty$$

Note that if $\lambda = 0$ then the density of X in (1.2.2) reduces to a normal distribution. The proof that equation (1.2.1) is a true density function comes from the following lemma

Lemma 1.2.1 *Azzalini [1985] Let f be a symmetric density function with respect to 0, G an absolutely continuous distribution function such that G' is symmetric with respect to 0, then*

$$2G(\lambda y)f(y) \quad (-\infty < y < \infty) \tag{1.2.3}$$

is a density function for any real λ.

Proof Let Y and X be independent random variables with density f and G', respectively. Then

$$1/2 = P\{X - \lambda Y < 0\} = E_Y(P\{X < \lambda y | Y = y\}) = \int_{-\infty}^{\infty} G(\lambda y) f(y)\, \mathrm{d}y \quad (1.2.4)$$

Another characterization of the skew-normal is as follow

Proposition 1.2.1 *Azzalini and Dalle-Valle [1996] If* Y_0 *and* Y_1 *are independent* $N(0,1)$ *variables and* $\xi \in (-1,1)$ *Then*

$$Z = \xi |Y_0| + \sqrt{1+\xi^2} Y_1 \;\; follows \;\; SN_1(\lambda(\xi)) \quad (1.2.5)$$

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

1.2.2 Multivariate skew-normal distribution

Extensions to the univariate skew-normal distribution in equation (1.2.1) was first proposed by Azzalini [1985] and expanded further by Azzalini and Dalle-Valle [1996]. Later on, many authors have worked on generalizing such findings to include a family of multivariate and skew-elliptical distributions. This section states some relevant results in chronological order, and concludes with two advancements and their proof.

Azzalini and Capitanio [1999] defined a multivariate skew-distribution by the following non-unique notation

$$f(z|Q) = 2 f_k(z) Q(z) \quad z \in \Re^k \quad (1.2.6)$$

where $f_k(.)$ is the density corresponding to a l-dimensional elliptical distribution, defined in Definition (1.2.1), and Q is a skewing function such that $Q(z) \geq 0$ and $Q(-z) = 1 - Q(z), \;\; \forall z \in \Re^k$. Note that $Q(.)$ could equally be represented by $Q(z) = \nu(u(z))$, for some function $u : \Re^k \to \Re$ and some non-negative function $\nu : \Re \to \Re$, such that $u(-z) = -u(z), \;\; \forall z \in \Re^k$, and $\nu(-u) = 1 - \nu(u), \;\; \forall u \in \Re$. Also note, equation (1.2.6) is a generalization to (1.2.1).

Definition 1.2.1 *Owen and Rabinovitch [1983] A random vector $X = (X_1, \dots, X_p)$ is said to have a p-elliptical distribution if it has a density function $f_X(x)$ say, then $f_X(x)$ can be expressed in the form of*

$$F_X(x) = k|\Omega|^{-1} g\Big((x-\mu)^T \Omega^{-1} (x-\mu)\Big)$$

for some function $g(.)$ mapping from non-negative reals to non-negative reals, and $g(.)$ is independent of k. Ω is a positive definite matrix , and μ is the mean vector. Thus $f_X(x)$ is only a function of the quadratic form $(x-\mu)^T\Omega^{-1}(x-\mu)$, which is positive by definition.

Some advancements are represented in the work of Arellano-Valle et al. [2002], where they generalized the previous findings in (1.2.6) to a class of skew-symmetric distributions starting with a family of special C-class symmetric distribution, where C represents the class of all symmetric random vectors X with $P(X = 0) = 0$ and $|X| = (|X_1|, \dots, |X_m|)^T$ and $sign(X) = (W_1, \dots, W_m)^T$ being independent, and $sign(X) \sim U_m \stackrel{d}{=}$ uniform on $\{-1,1\}^m$, such that

$$W_i = \begin{cases} +1 & \text{if } X_i > 0 \\ -1 & \text{if } X_i \leq 0 \end{cases} \quad i = 1, \ldots, m, \quad , U_m \sim \text{ uniform on } \{-1,1\}^m. \tag{1.2.7}$$

Hence to obtain a density of any arbitrary skew distribution, the following notations hold

$$f(z|a_m) = {K_m}^{-1} f_k(z) Q_m(z), \quad \forall z \in \Re^k \tag{1.2.8}$$

where

$$K_m = P(X > 0) \quad \text{ and } \quad Q_m(z) = P(X > 0|Z = z) \tag{1.2.9}$$

for some random vectors X and Z with dimensions $m \times 1$ and $k \times 1$ respectively, and with joint distribution, in which that Z has marginal density f_k. Note that if X is a C-random vector then $K_m = P(X > 0) = 2^{-m}$ which transforms (1.2.8) to

$$f(z|a_m) = 2^m f_k(z) Q_m(z), \quad \forall z \in \Re^k \tag{1.2.10}$$

Finally, for convenience the following results are used in the later sections. A modified version of Arellano-Valle and Genton [2005] states that

Definition 1.2.2 *An n-dimensional random vector X follows a skew-normal distribution with location vector $\mu \in \Re^n$, dispersion matrix Σ (a $n \times n$ positive definite matrix) and a skewness vector $\lambda \in \Re^n$, if its pdf is given by*

$$sn_n(x|\mu, \Sigma, \lambda) = 2\phi_n(x|\mu, \Sigma)\Phi_1(\lambda^T \Sigma^{-1/2}(x - \mu)), \quad x \in \Re^n. \tag{1.2.11}$$

Remark 1.2.1 *Note that since the condition that $\Phi_1(-w) = 1 - \Phi_1(w)$ for all*

$w \in \Re$ satisfies requirement (1.2.6) and hence is sufficient to guarantee that (1.2.11) is a pdf.

Azzalini and Dalle-Valle [1996] proposed a simplified parametrization to $\Phi_1(.)$ in (1.2.11) of λ in terms of an arbitrary $n \times n$ positive definite matrix Δ. Let us say $\Delta = \Sigma$ or $\Delta = I_n$, such that $\delta^T \Delta^{-1} \delta < 1$ for some $\delta \in \Re^n$ then,

$$\lambda = \frac{\Delta^{1/2}\delta}{\sqrt{1 - \delta^T \Delta^{-1} \delta}} \tag{1.2.12}$$

Arellano-Valle and Genton [2005] representation of the multivariate skew-normal distribution is basically a modification to (1.2.5) in Proposition (1.2.1), combined with (1.2.12) yielding

Proposition 1.2.2 *Arellano-Valle et al. [2005] Let $W \sim SN_n(\lambda)$. Then*

$$W \stackrel{d}{=} \delta|X_0| + (I_n - \delta\delta^T)^{1/2} X_1, \quad \text{where} \quad \delta = \frac{\lambda}{\sqrt{1 + \lambda^T \lambda}} \tag{1.2.13}$$

$X_0 \sim N_1(0, 1)$ independent of $X_1 \sim N_n(0, 1)$

Before proving the previous proposition the following lemma is needed.

Lemma 1.2.2 *Arellano-Valle et al. [2005] Let $Y \sim N_p(\mu, \Sigma)$ and $X \sim N_q(\nu, \Omega)$ Then*

$$\begin{aligned} \phi_p(y|\mu + Ax, \Sigma)\phi_q(x|\nu, \Omega) &= \phi_p(y|\mu + A\nu, \Sigma + A\Omega A^T) \\ &\quad \times \phi_q(x|\nu + \Lambda A^T \Sigma^{-1}(y - \mu - A\nu), \Lambda) \end{aligned} \tag{1.2.14}$$

where $\Lambda = (\Omega^{-1} + A^T \Sigma^{-1} A)^{-1}$.

Proof of Lemma (1.2.2) By letting $z = y - \mu - A\nu$ and $w = x - \nu$, we have after some standard algebraic operations

$$(z - Aw)^T\Sigma^{-1}(z - Aw) + w^T\Omega^{-1}w = z(\Sigma + A\Omega A^T)^{-1}z \\ + (w - \Lambda A^T\Sigma^{-1}z)^T\Lambda^{-1}(w - \Lambda A^T\Sigma^{-1}z),$$

Noting that $|\Sigma + A\Omega A^T||\Lambda| = |\Sigma||\Omega|$.

Proof of Proposition (1.2.2) Let $W = \delta|X_0| + (I_n - \delta\delta^T)^{1/2}X_1$. Since $W||X_0| \sim N_n(\delta|X_0|, I_n - \delta\delta^T)$ where $|X_0| \sim HN(0,1)$ then by Lemma (1.2.2) it follows that

$$f_W(w) = \int_0^\infty 2\phi_n(w|\delta t, I_n - \delta\delta^T)\phi_1(t)\mathrm{d}t = \int_0^\infty 2\phi_n(w|0, I_n)\phi_1(t|\delta^T w, 1 - \delta^T\delta)\mathrm{d}t \\ = 2\phi_n(w)\Phi_1(\frac{\delta^T w}{\sqrt{1 - \delta^T\delta}})$$

Then $W \sim SN_n(\lambda)$ with $\lambda = \frac{\delta}{\sqrt{1-\delta^T\delta}}$.

The following is another needed and useful proposition.

Proposition 1.2.3 *Arellano-Valle et al. [2005] Suppose that $Y|T = t \sim N_n(\mu + dt, \Psi)$ and $T \sim HN_1(0,1)$ (a standardized half-normal distribution). Let $\Sigma = \Psi + dd^T$. Then the joint distribution of $(Y^T, T)^T$ can be written as*

$$f_{Y,T}(y, t|\theta, \lambda) = 2\phi_n(y|\mu, \Sigma)\phi_1(t|\nu, \tau^2)\mathbf{I}\{t > 0\} \tag{1.2.15}$$

where

$$\nu = \frac{d^T\Psi^{-1}(y - \mu)}{1 + d^T\Psi^{-1}d} \;\; and \;\; \tau^2 = \frac{1}{1 + d^T\Psi^{-1}d} \tag{1.2.16}$$

Proof The proof for Proposition (1.2.3) follows directly from knowing that the joint density of Y and T is

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

and from algebraic manipulation by integrating t out we have

$$\phi_n(y|\mu + dt, \Psi) = \phi_n(y|\mu, \Sigma)\Phi_1(t|\nu, \tau^2)$$

which concludes the proof. Consequently, the marginal distribution of Y of (1.2.15) is given by

$$f_Y(y|\theta, \lambda) = 2\phi_n(y|\mu, \Sigma)\Phi_1(\frac{\nu}{\tau})$$

So far, the literature have presented multiple useful versions of constructing an n-variate skew-normal distribution. Hence, to complete the set of findings and to draw some useful characteristic of this distribution, the following corollary gives the expectation and variance of $SN_n(.)$ and follows directly from previous results.

Corollary 1.2.1 *Arellano-Valle et al. [2005] Let* $Y \overset{d}{=} \mu + \Sigma^{1/2}W$*, where,* $W \sim SN_n(\lambda)$*. Then* $Y \sim SN_N(\mu, \Sigma, \lambda)$*. Moreover,*

$$E[Y] = \mu + \sqrt{\frac{2}{\pi}}\Sigma^{1/2}\delta \;\; and \;\; Var[Y] = \Sigma - \frac{2}{\pi}\Sigma^{1/2}\delta\delta^T\Sigma^{1/2}.$$

Chapter 2

Modeling the joint distribution via a skew-normal copula

2.1 The model

Instead of building an additive model similar to one presented in (1.1.1) one can consider a general multivariate distribution(MV) of the form

$$Y_i|b_i \sim MV(\eta(X_i\beta + b_i), \Sigma_i(\phi_i, t_i)) \tag{2.1.1}$$

where $Y_i|b_i$ is the response variable of the i^{th} unit at time $t_i = (t_{i1}, \ldots, t_{in_i})$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$, the longitudinal data available for unit i is then $Y_i = (Y_{i1}, \ldots, Y_{ij}, \ldots, Y_{in_i})^T$. Further, x_i is a $(n_i \times p)$ vector of explanatory variables for unit i at time t_i, b_i is the unobserved unit specific random effect and β is a $(p \times 1)$ vector of regression coefficients to be estimated.

Assume the marginal densities $Y_{ij}|b_i$ is a function of x_{ij} , t_{ij} , b_i and β

via the same link function $\eta(.)$. Therefore, let $F_{Y_{ij}}(y_{ij}|x_{ij}, t_{ij}, b_i, \beta)$ denote the conditional marginal distribution function of the response variable $Y_{ij}|b_i$, and $f_{Y_{ij}}(y_{ij}|x_{ij}, t_{ij}, b_i, \beta)$ denote its density. We may assume for the sake of simplicity that x_{ij} does not depend on t_i or b_i. Further, let $F_b(b_i|\sigma_b^2) \sim N_1(0, \sigma_b^2)$ be the distribution of the unit specific random effect with density $f_b(.|\sigma_b^2)$.

2.2 Skew-normal copula

After characterizing the joint and marginal conditional distributions of $Y_i|b_i$ in the previous section, this section defines the general properties of a copula and more specifically construct a skew-normal copula with the given marginal distributions $Y_{ij}|b_i$.

Definition 2.2.1 Nelsen [1999] *$C : [0,1]^d \to [0,1]$ is a d-dimensional copula if C is a joint cumulative distribution function of a d-dimensional random vector on the unit cube $[0,1]^d$ with uniform marginals.*

For example, in a bivariate case $C : [0,1] \times [0,1] \to [0,1]$ is a bivariate copula if $C(0,u) = C(u,0) = 0$, $C(1,u) = C(u,1) = u$ and $C(y_1,y_2) - C(x_1,y_2) - C(y_1,x_2) + C(x_1,x_2) \geq 0$ for all $[x_1,y_1] \times [x_2,y_2] \subseteq [0,1] \times [0,1]$.

Definition 2.2.2 Dodge [2003] *Suppose that a random variable X has a continuous distribution for which the cumulative distribution function is F_X. Then the random variable W defined as*

$$W = F_X(x)$$

has a uniform distribution.

Considering definitions (2.2.1) and (2.2.2), a skew-normal copula specifies the joint distribution of $Y_i|b_i$ with specified marginal distributions $F_{Y_{ij}}(y_{ij}|.)$ as follow, let

$$Z_{ij} \sim SN_1(b_i, 1, \lambda_{ij})$$

at time t_{ij} for unit i, then

$$Z_i = (z_{i1}, \ldots, z_{ij}, \ldots, z_{in_i})^T \sim SN_{n_i}(b_i 1, \Sigma_i, \lambda_i)$$

where Σ_i is a correlation matrix, which has all its diagonal elements equal to 1, and $1 = (1, \ldots, 1)^T$. For modeling the contribution of $Y_i|b_i$, the results obtained in (1.2.11) allow us to utilize the standard formation of a copula in Definition (2.2.1) and define a multivariate skew-normal copula as

$$C_i^{b_i} = C_i^{b_i}(u_{i1}, \ldots, u_{n_i}) = SN_{n_i}(SN_1^{-1}(u_{i1}), \ldots, SN_1^{-1}(u_{in_i})|b_i 1, \Sigma_i, \lambda_i) \quad (2.2.1)$$

where $u_{ij} \sim \text{uniform}(0, 1)$.

From Definition (2.2.2) we have $F_{Y_{ij}}(y_{ij}|.) \sim u_{ij}$ similarly $SN_1(z_{ij}|.) \sim u_{ij}$, therefore we can write

$$\begin{aligned} Y_{ij} &= F_{Y_{ij}}^{-1}(u_{ij}|.) = F_{Y_{ij}}^{-1}(SN_1(z_{ij})|.) \\ Z_{ij} &= SN_1^{-1}(u_{ij}) = SN_1^{-1}(F_{Y_{ij}}(y_{ij}|.)) \end{aligned} \quad (2.2.2)$$

Therefore, the joint distribution function of $Y_i|b_i$ is modeled as

$$F_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,b_i,\Sigma_i) = SN_{n_i}(SN_1^{-1}(F_{Y_{i1}}(y_{i1}|x_{i1},\beta,b_i)),\ldots \\ ,SN_1^{-1}(F_{Y_{in_i}}(y_{in_i}|x_{in_i},\beta,b_i)|b_i 1,\Sigma_i,\lambda_i) \tag{2.2.3}$$

The corresponding skew-normal copula density function conditioned on b_i is

$$\begin{aligned} f_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,b_i,\Sigma_i,) &= \prod_{j=1}^{n_i} \frac{\partial F_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,b_i,\Sigma_i)}{\partial y_{ij}} \\ &= \frac{\partial SN_{n_i}(SN_1^{-1}(F_{Y_{i1}}(y_{i1}|.)),\ldots,SN_1^{-1}(F_{Y_{in_i}}(y_{in_i}|.)|b_i 1,\Sigma_i,\lambda_i)}{\partial y_{i1}\ldots\partial y_{in_i}} \\ &= sn_{n_i}(z_{1i},\ldots,z_{in_i}|b_i 1,\Sigma_i,\lambda_i)\prod_{j=1}^{n_i}\frac{\partial SN_1^{-1}(F_{Y_{ij}}(y_{ij}|.))}{\partial y_{ij}} \\ &= sn_{n_i}(z_{1i},\ldots,z_{in_i}|b_i 1,\Sigma_i,\lambda_i)\prod_{j=1}^{n_i}\frac{f_{Y_{ij}}(y_{ij}|x_{ij},\beta,b_i)}{sn_1(z_{ij}|b_i,1,\lambda_{ij})} \end{aligned} \tag{2.2.4}$$

where $z_{ij} = SN_1^{-1}(F_{Y_{ij}}(y_{ij}|.)|b_1,1,\lambda_{ij})$, the density function becomes

$$f_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,b_i,\Sigma_i,) = sn_{n_i}(z_{1i},\ldots,z_{in_i}|b_i 1,\Sigma_i,\lambda_i)\prod_{j=1}^{n_i}\frac{f_{Y_{ij}}(y_{ij}|x_{ij},\beta,b_i)}{sn_1(z_{ij}|b_i,1,\lambda_{ij})} \tag{2.2.5}$$

The unconditional skew-normal copula density function is

$$f_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,\Sigma_i) = \int f_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,b_i,\Sigma_i)f_{b_i}(0,\sigma_b^2)\mathrm{d}b_i \tag{2.2.6}$$

where $f_b(0,\sigma_b^2) \sim N_1(0,\sigma_b^2)$.

Note that the joint distribution of the response vector $Y_i|b_i$ has not yet been explicitly defined, all what is required so far is to explicitly define that the marginal distributions $f_{Y_{ij}}(y_{ij}|.)$. These results demonstrate that one can choose a copula representing the joint distribution function of any multivariate distribution by having all the marginals distribution functions specified. For this thesis we define the marginals as as a member of the exponential family distributions, as explained later in section (2.4). The next section proposes a modification to the correlation matrix Σ_i to include the time dependence variable. Later we propose a GLM framework for the response variable.

To conclude this section, the following summarize our model. For $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$ our model is of the form

$$Y_i|b_i \sim F_{Y_i}(.|x_i, \beta, b_i, t_i, \Sigma_i), \quad Y_{ij}|b_i \sim F_{Y_{ij}}(.|x_{ij}, \beta, b_i, t_{ij})$$

$$F_{Y_i}(.|x_i, \beta, b_i, t_i) = SN_{n_i}(Z_i|b_i 1, \Sigma_i, \lambda_i), \text{where} \quad Z_i = (Z_{i1}, \ldots, Z_{in_i})^T$$

$$Z_{ij} \sim SN_1(b_i, 1, \lambda_{ij}), \quad Y_{ij} = F_{ij}^{-1}(SN_1(z_{ij}|b_i, 1, \lambda_{ij})|x_i, \beta, b_i, t_i, \Sigma_i), \quad b_i \sim N_1(.|\sigma_b^2)$$

To facilitate comparison to other similar models, the model of Arellano-Valle et al. [2005] is stated as

$$\mathbf{Y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}b_i + \epsilon_\mathbf{i} \tag{2.2.7}$$

$$b_i \overset{iid}{\sim} SN_q(0, D, \lambda_b), \quad \epsilon_i \overset{iid}{\sim} SN_{n_i}(0, \psi_i, \lambda_{e_i}), \quad i = 1, \ldots, m$$

where b_i is independent of ϵ_i and,

$$Y_i|b_i \overset{ind}{\sim} SN_{n_i}(\mathbf{X_i}\beta + \mathbf{Z_i}b_i, \psi_i, \lambda_{e_i}) \tag{2.2.8}$$

2.3 Autoregressive correlation matrix

To produce a more plausible model one needs to take into account the different sources of random variation within observations. Such random variations could be generally specified by the following.

(i) **Random effects**: by sampling at random from a population, multiple aspects of behavior could persist stochastically. Such is the case when the average level of the response variable varies widely between different time intervals. At some moments there could be a very low response, others could have a higher response. Such behaviors could produce some bias within the error.

(ii) **Serial correlation**: different observations from the same variable sampled at different time intervals could be highly correlated. For example, a patient's blood pressure measure repeatedly over time in a medical study. Therefore, one needs to take into account a regressive correlation structure of the error.

(iii) **Measurement error**: a general random error could be accounted for between observations. When two measurements are taken simultaneously from a source, there could be a slight variation between them which causes a random error.

Note that the random effect source of variation is already accounted for within the model as a random intercept b_i in (3.2.6). Therefore, we would only consider integrating the serial correlation and the measurement error effects. The variance-covariance matrix Σ_i presented in (2.1.1) could be modeled as a function of time and a dispersion variable ϕ_i as

$$\Sigma_i = \Sigma_i(\phi_i, t_i) \tag{2.3.1}$$

As a result, assuming an additive decomposition of the above sources in the variance-covariance structure letting

$$Y_{ij} = W_i(y_{i1}, \ldots, y_{i(j-1)}, t_{ij}) + \epsilon_{ij} \tag{2.3.2}$$

where $W_i(.)$ is a linear function composing the source of variation generated from the serial correlation to previous observations. Here, let $\epsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2_{\epsilon_i})$ be sample of n_i independent copies of a stationary Gaussian process with mean zero and variance $\sigma^2_{\epsilon_i}$ that represents the measurement error.

Since the response variables is of the form $Y_i|b_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in_i})$, define, H_i as $n_i \times n_i$ matrix with $(j,k)^{\text{th}}$ element $h_{ijk} = \rho_i(|t_{ij} - t_{ik}|)$, h_{ijk} is the correlation between Y_{ij} and Y_{ik} represented in the function $W_i(.)$ above. Also, let I_i be $n_i \times n_i$ identity matrix. Since the marginal distribution of $Y_i|b_i$ in (2.2.2) are assumed to have the same distribution family, the variance-covariance matrix of $Y_i|b_i$ is easily constructed as

$$Var(Y_i) = \sigma_i^2 H_i + \sigma^2_{\epsilon_i} I_i \tag{2.3.3}$$

Now, σ_i^2 is the variance of each y_{ij} and the correlations amongst the Y_i's are determined by the autocorrelation function $\rho_i(.)$ as

$$Cov(Y_{ij}, Y_{ik}) = \sigma_i^2 \rho_i(|t_{ij} - t_{ik}|) \tag{2.3.4}$$

The simplest form to express the serial correlation above is to assume an explicit dependence of the current Y_{ij} on predecessors $Y_{i(j-1)}, \ldots, Y_{i1}$, which could be modeled using n^{th} order autoregressive model. For example, considering a first order autoregressive model as

$$y_{ij} = \alpha_i y_{i(j-1)} + \epsilon_{ij} \tag{2.3.5}$$

where ϵ_{ij} as defined in (2.3.2)

Note that it would be difficult to give a closed interpretation of the α parameter if the measurements are not equally spaced in time or when times of measurements are not common to all units. One way of solving this issue to implement an exponential autocorrelation function $\rho(.)$, where

$$Cov(Y_{ij}, Y_{ik}) = \sigma_i^2 e^{-\phi_i |t_{ij} - t_{ik}|} \tag{2.3.6}$$

Furthermore, the correlation between two response variable becomes

$$Corr(Y_{ij}, Y_{ik}) = e^{-\phi_i |t_{ij} - t_{ik}|} \tag{2.3.7}$$

This correlation structure would later be used to construct a correlation coefficient matrix $\Sigma_i = \Sigma_i(\phi_i, t_i)$ that could represents the bivariate relation between marginal response variables $Y_{ij}|b_i$, henceforth, implemented in the copula structure.

2.4 GLM framework for response variable and the maximum likelihood

We have already defined implicitly the joint distribution of the response variable $Y_i|b_i$. Moreover, we have modeled such joint distribution using a skew-normal copula by implementing a fist order autoregressive correlation structure using an exponential autocorrelation function. For the sake of completeness and rigor, at this stage we need to explicitly characterize the expression of the distribution of $Y_i|b_i$ defined in (2.1.1). To allow flexibility, this thesis establishes a general expression for the distribution being a member of the exponential family of distributions. Later on, both the exponential and gamma distributions are used in modeling and simulation with different non-linear link functions.

2.4.1 Exponential family of distributions

Consider an over-dispersed exponential family of distributions, which is a generalization of the exponential family and exponential dispersion model of distributions. It includes those probability distributions parameterized by θ and τ, whose density functions can be expressed in the form

$$f_Y(y|\theta,\tau) = \exp\left\{\frac{b(\theta)^T T(y) - A(\theta)}{d(\tau)} + B(y,\tau)\right\} \tag{2.4.1}$$

where $T(.), b(.), d(.), A(.)$ and $B(.)$ are known functions and θ is the parameter relating to the mean of the model, τ is the dispersion parameter.

Furthermore, we consider a link function $\eta(.)$ that relates to the mean of the model by

$$E(Y|b_i) = \eta(X\beta + b_i)$$

The above general distribution structure include many known distributions; for example, Exponential, Gamma, Pareto, and Poisson distribution.

2.4.2 Response variable as an exponential distribution

By assuming the marginals of the response vector $Y_i|b_i$ follow an exponential distribution as

$$\begin{aligned} f_{Y_{ij}}(y_{ij}|\lambda) &= \exp\{-\lambda y_{ij} + \log(\lambda)\}, \quad y_{ij} \geq 0 \\ A(.) &= \log(\lambda) \\ B(.) &= 0 \\ E(Y_{ij}) &= \mu = \frac{1}{\lambda} = \eta(x_{ij}\beta + b_i) \\ y_{ij} \geq 0 &\Rightarrow \mu = \eta(x_{ij}\beta + b_i) \geq 0 \end{aligned} \tag{2.4.2}$$

then

$$f_{Y_{ij}}(y_{ij}|x_{ij}, b_i, \beta) = \exp\{\frac{-y_{ij}}{\eta(x_{ij}\beta + b_i)} - \log(\eta(x_{ij}\beta + b_i))\}, \quad y_{ij} \geq 0 \tag{2.4.3}$$

where log is the natural logarithm.

2.4.3 Response variable as a gamma distribution

Similarly, by assuming the marginals of the response vector $Y_i|b_i$ follow gamma distribution as

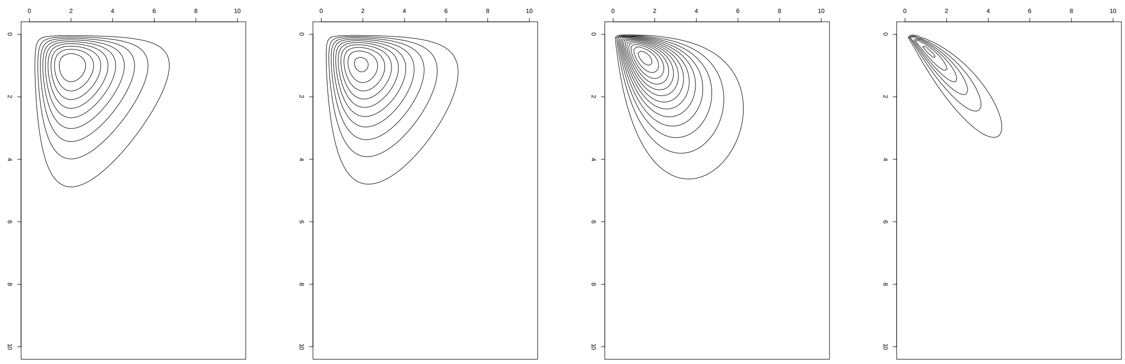
$$\begin{aligned}
f_{Y_{ij}}(y_{ij}|\alpha,k) &= \exp\{-\frac{y_{ij}}{\alpha} - A(\alpha,k) + B(y_{ij},k)\} \quad y_{ij} \geq 0 \\
B(y_{ij}) &= (k-1)\log(y_{ij}) \\
A(\alpha,k) &= k\log(\alpha) + \log(\Gamma(k)) \\
\Gamma(k) &= (k-1)! \quad k \geq 0 \\
E(Y_{ij}) &= \mu = k\alpha = \eta(x_{ij}\beta + b_i) \\
y_{ij} \geq 0, \quad k &\geq 0 \Rightarrow \mu = \eta(x_{ij}\beta + b_i) \geq 0
\end{aligned} \tag{2.4.4}$$

then

$$\begin{aligned}
f_{Y_{ij}}(y_{ij}|x_{ij},b_i,\beta) = \exp\Bigg\{ &- \frac{k y_{ij}}{\eta(x_{ij}\beta + b_i)} - k\log\left(\frac{\eta(x_{ij}\beta + b_i)}{k}\right) \\
&- \log(\Gamma(k)) + (k-1)\log(y_{ij})\Bigg\}
\end{aligned} \tag{2.4.5}$$

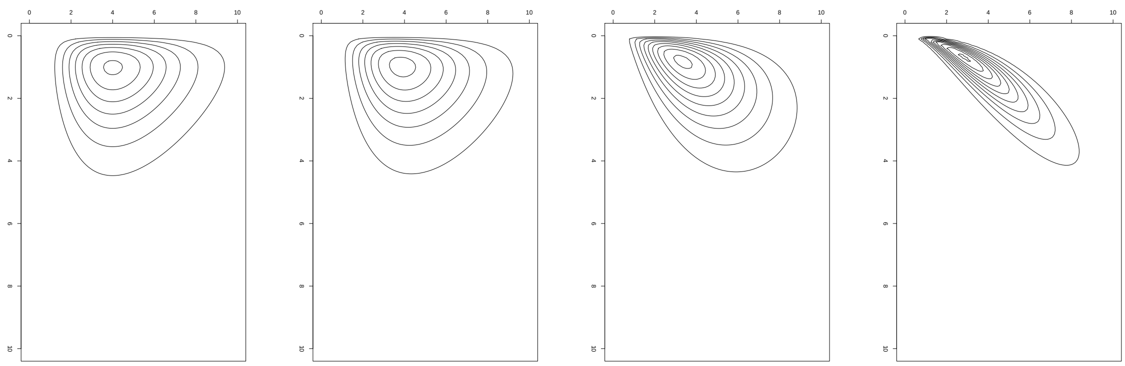
2.5 Graphical examples

This section presents a variety of bivariate contour plots to compare the approximation of the method proposed in section (2.2). Mainly, a gamma density function is used under multiple configuration of shape parameter k and correlation ρ. In figures (2.1), (2.2) and (2.3) when $\rho = 0$, the contour plot is generated using two independent gamma densities; when the $\rho \neq 0$ the contour is generated using the copula method in (2.2) and (2.2.5).



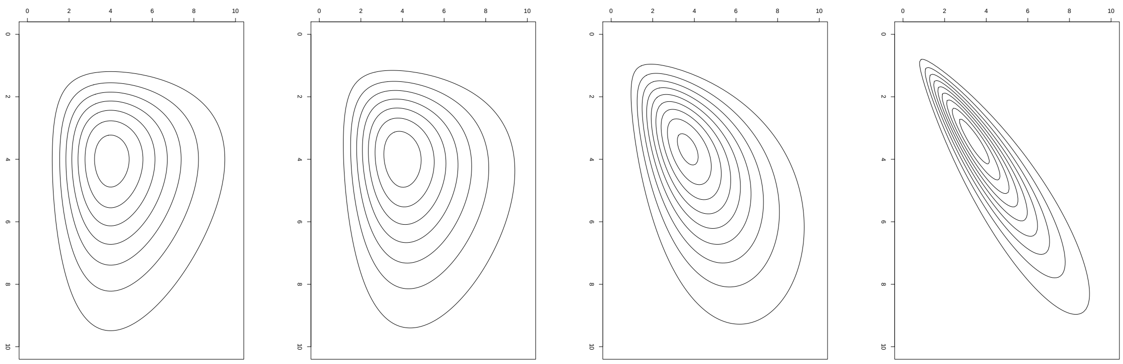
(a) For $\rho = 0$ (b) For $\rho = 0.1$ (c) For $\rho = 0.5$ (d) For $\rho = 0.9$

Figure 2.1: Contour of a bivariate gamma distribution using the proposed copula method in (2.2) and (2.2.5) with shape parameter $k = (2, 3)^T$, scale $\alpha = (1, 1)^T$ and different values of correlation coefficient ρ.



(a) For $\rho = 0$ (b) For $\rho = 0.1$ (c) For $\rho = 0.5$ (d) For $\rho = 0.9$

Figure 2.2: Contour of a bivariate gamma distribution using the proposed copula method in (2.2) and (2.2.5) with shape parameter $k = (2, 5)^T$, scale $\alpha = (1, 1)^T$ and different values of correlation coefficient ρ.



(a) For $\rho = 0$ (b) For $\rho = 0.1$ (c) For $\rho = 0.5$ (d) For $\rho = 0.9$

Figure 2.3: Contour of a bivariate gamma distribution using the proposed copula method in (2.2) and (2.2.5) with shape parameter $k = (5, 5)^T$, scale $\alpha = (1, 1)^T$ and different value of correlation coefficient ρ.

The similarities between figures produced using the two methods are very apparent. Moreover, the application of copulas allows more flexibility, where any correlation structure could easily be modeled even when the joint distribution is not explicitly defined.

2.6 Likelihood function

The setting of longitudinal data over several units with repeated observations has an additive characteristic of the log-likelihood across units. In particular, given m units as $i = 1, \ldots, m$ and the model setting in (2.1.1) , we have the unconditional likelihood function as

$$\begin{aligned}
\mathcal{L}(\beta,\Sigma) &= \prod_{i=1}^{m}\int f_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,\Sigma_i,\lambda_i,b_i)f_{b_i}(0,\sigma_b^2)\mathrm{d}b_i \\
&= \int\prod_{i=1}^{m}\left\{f_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,\Sigma_i,\lambda_i,b_i)\right\}f_{b_i}(0,\sigma_b^2)\mathrm{d}b_i \\
&= \int\prod_{i=1}^{m}\left\{sn_{n_i}(z_{1i},\ldots,z_{in_i}|b_i1,\Sigma_i,\lambda_i)\prod_{j=1}^{n_i}\frac{f_{Y_{ij}}(y_{ij}|x_{ij},\beta,b_i)}{sn_1(z_{ij}|b_i,1,\lambda_{ij})}\right\}f_{b_i}(0,\sigma_b^2)\mathrm{d}b_i \\
&= \int\prod_{i=1}^{m}\mathcal{L}_i(\beta,\lambda_i,\sigma_b,\Sigma_i|b_i)f_{b_i}(0,\sigma_b^2)\mathrm{d}b_i
\end{aligned} \tag{2.6.1}$$

where $\mathcal{L}_i(\beta,\lambda_i,\sigma_b,\Sigma_i|b_i)$ is the conditional likelihood. The exchangeability of the product and the integral comes as a result of the independent assumption between units, similar to the independent assumption of errors in cross-sectional data.

The complete conditional log-likelihood transforms to

$$\begin{aligned}
\ell(\beta,\lambda,\sigma_b,\Sigma|b_i) &= \log\left\{\prod_{i=1}^{m}\mathcal{L}_i(\beta,\lambda_i,\sigma_b,\Sigma_i|b_i)f_{b_i}(0,\sigma_b^2)\right\} \\
&= \sum_{i=1}^{m}\log\left\{f_{Y_i}(y_{i1},\ldots,y_{in_i}|x_i,\beta,\Sigma_i,\lambda_i,b_i)f_{b_i}(0,\sigma_b^2)\right\} \\
&= \sum_{i=1}^{m}\log\left\{sn_{n_i}(z_{1i},\ldots,z_{in_i}|b_i1,\Sigma_i,\lambda_i)\prod_{j=1}^{n_i}\frac{f_{Y_{ij}}(y_{ij}|x_{ij},\beta,b_i)}{sn_1(z_{ij}|b_i,1,\lambda_{ij})}f_{b_i}(0,\sigma_b^2)\right\} \\
&= \sum_{i=1}^{m}\ell_i(\beta,\lambda_i,\sigma_b,\Sigma_i|b_i)
\end{aligned} \tag{2.6.2}$$

where from (2.2.5) we have

$$\begin{aligned}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|b_i) &= \log(f_{Y_i}(y_{i1}, \ldots y_{in_i}|x_i, \beta, \Sigma_i, \lambda_i, b_i)) \\ &+ \log(f_{b_i}(0, \sigma_b)) \\ &= \log(sn_{n_i}(z_{i1}, \ldots, z_{in_i}|b_i 1, \Sigma_i, \lambda_i)) + \sum_{j=1}^{n_i} \log(f_{Y_{ij}}(y_{ij}|b_i, \beta, x_{ij})) \\ &- \sum_{j=1}^{n_i} \log(sn_1(z_{ij}|b_i, 1, \lambda_{ij})) + \log(f_{b_i}(0, \sigma_b))\end{aligned} \tag{2.6.3}$$

Note that we have delayed characterizing completely the log-likelihood function above to Chapter (3), where we introduce the MC-EM algorithm. Chapter (3) also specifies the exact form of skew-normal density functions implemented in simulation, as well as the exact form of marginal densities $Y_{ij}|b_i$.

Chapter 3

Numerical computation with EM-algorithm

3.1 Monte Carlo based EM algorithm

The expectation-maximization algorithm is an iterative method for maximizing the likelihood function. It consists of two steps from which the naming is derived; an expectation (E-step) and a maximization step (M-step). The E-step computes the expectation of the log-likelihood function over the given set of parameters $\theta \in \Theta$ and an unobserved parameter u with assumed density function $g(u|x, \theta)$. The M-step computes a new set of parameters that maximize the expected value of the log-likelihood function found earlier in the E-step. Those two procedures alternate on the path to find a set of parameters that maximize the likelihood function. Furthermore, let θ donate the model parameters when complete-data is available; hence, $\ell(\theta),\ \theta \in \Theta$ donates the complete-data log-likelihood function and $Q(\theta|\theta')$ the expected complete-data log-likelihood. Therefore both steps are as follow

- E-step: Computes $Q(\theta|\theta^{(r)}) = E_{u|\theta^{(r)}}[\ell(\theta)|\theta^{(r)}]$ as a function of θ.
- M-step: Find $\theta^{(r+1)}$ such that $Q(\theta^{(r+1)}|\theta^{(r)}) = max_{\theta\in\Theta}Q(\theta|\theta^{(r)})$.

Theoretically, the log-likelihood function $\ell(.)$ under the found parameter set $\theta^{(r)}$ should converge to a local or global maximum.

The EM-algorithm has been used widely in literature under different naming conventions. Nevertheless, one of the earliest explanations of such method was published by Dempster et al. [1977], where they generalized earlier attempts and sketched a convergence analysis for a wider class of problems. Since then, numerous uses of such method have been unified under the name of EM-algorithm.

An advancement of Meng and Rubin [1993] studies computational difficulties encountered while computing the M-step, where they proposed smaller maximization steps over the parameter space. They argued that instead of maximizing the whole set of parameters one can maximize in a sequential manner a subset of parameters independently, while the other subset is held fixed. Such modification is called a constrained maximization step (CM). Theoretically, as long as the maximization is applied on the whole set of parameters the algorithm should reach convergence. This modification is referred to as an ECM-algorithm.

A second important advancement to the EM-algorithm was proposed by Wei and Tanner [1990], and is called the Monte Carlo EM algorithm. By applying the law of large numbers on the E-step above, one can approximate $Q(\theta|\theta^{(r)})$ as

$$\begin{aligned} Q(\theta|\theta^{(r)}) = E[\ell(\theta)|\theta^{(r)}] &= \int \ell(\theta|x,u)g(u|x,\theta^{(r)})\mathrm{d}u \\ &\cong \frac{1}{R}\sum_{t=1}^{R}\ell(\theta^{(r)}|u^{(t)}) \end{aligned} \tag{3.1.1}$$

where u is the unobserved variable and R is relatively a large sample size.

It is important to mention the work of Wu [1983] which studied the convergence properties of the EM algorithm. Wu [1983]'s work elaborates on Dempster et al. [1977] by clearly indicating a set of conditions that govern the convergence of EM-algorithm to a stationary point, whether a global or a local maximum. Some of those conditions are; insuring that $Q(\theta|\theta')$ defined (3.1.1) is continuous on both θ and θ'; the set of parameters to be estimated θ' belong to a compact space, let's say Ω_0; and the log-likelihood is bounded. The bound condition of the proposed log-likelihood depends severely on the initial starting point of the EM-algorithm, where it involves terms like $log(|\Sigma_i|) \to -\infty$ for very small $|\Sigma_i|$, shown later in equation (3.2.13), where $|.|$ is the determinant. Heuristic method of initiating the algorithm from different starting points was successfully used by Arellano-Valle et al. [2005], and considering the similarity of our model to theirs, a similar approach has been used here. Nevertheless, more analytical research has to be undergone in order to prove theoretically the compactness of Ω_0 and the continuity of $Q(\theta|\theta')$, especially for the dispersion parameter ϕ_i in (2.3.1), since it is shown to be convoluted in the term $|\Sigma_i|$. This thesis would mainly focus on implementing the EM-algorithm heuristically with hope to show more theoretical convergence characteristics in the future. The following two sections offer an elaborate explanation of the application of those two advancements to the proposed model above.

3.2 Applying the MC E-step

Let $\theta^{(r)} = (\beta^{(r)}, \Sigma_i^{(r)}, \sigma_b^{(r)}, \lambda_i^{(r)})$ be the parameters of the r^{th} EM iteration, then the E-step for unit i at $(r+1)$ EM iteration is

$$\begin{aligned} Q_i(\theta|\theta^{(r)}) &= E_{b_i|z_i}[\ell_i(\theta|x_i, b_i, y_i)|\theta^{(r)}] \\ &= \int \ell_i(\theta|x_i, y_i, b_i) f_{b_i|z_i}(b_i|z_i, \theta^{(r)}) \mathrm{d}b_i \\ &\cong \frac{1}{R_i} \sum_{j=1}^{R_i} \ell_i(\theta|x_i, y_i, b_i^{(j)}) \end{aligned} \tag{3.2.1}$$

where $b_i^{(j)}$ is the j^{th} sample generated from the distribution of $b_i|z_i, \theta^{(r)}$, R_i donates the number of replication on the i^{th} unit. Hence,

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{m} Q_i(\theta|\theta^{(r)}) \tag{3.2.2}$$

So far we haven't explicitly defined the distribution of $b_i|Z_i$. Nevertheless, we have mapped the response variables $Y_i|b_i$ to a skew-normal copula in Section (2.2) using a standard skew-normal marginal distributions as in equation (2.2.1) via the link in (2.2.2). Similarly, we can use a conditional skew-normal distributional as $Z_{ij}|b_i \sim SN_1(b_i, 1, \lambda_{ij})$, which by Azzalini [1985] and definition (1.2.2), for each j, has its pdf in the following form

$$f_{Z_{ij}|b_i}(z_{ij}|b_i) = 2\phi_1(z_{ij}|b_i, 1)\Phi_1(\lambda_{ij}(z_{ij} - b_i)) \tag{3.2.3}$$

Alternatively, by Proposition (1.2.3), we can write for any given v_i

$$Z_{ij}|v_i, b_i \sim N_1(b_i + \delta_{ij} v_i, 1 - \delta_{ij}^2) \tag{3.2.4}$$

$$v_i \sim HN_1(0,1) \quad b_i \sim N_1(0, \sigma_b^2)$$

where

$$\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$$

Similarly, in a multivariate case, take $Z_i|b_i \sim SN_{n_i}(b_i I, \Sigma_i, \lambda_i)$, where Σ_i has all its diagonal elements as 1, we have

$$f_{Z_i|b_i}(z_i, b_i|, \Sigma_i, \lambda_i) = 2\phi_{n_i}(z_i|b_i 1, \Sigma_i)\Phi_1(\lambda^T \Sigma_i^{-1/2}(z_i - b_i 1)) \tag{3.2.5}$$

and arguing as earlier, we write

$$Z_i|v_i, b_i \sim N_{n_i}(b_i 1 + \Sigma_i^{1/2}\delta_i^* v_i, \Sigma_i^{1/2}(I - \delta_i^* \delta_i^{*T})\Sigma_i^{1/2}) \tag{3.2.6}$$

$$v_i \sim HN_1(0,1), \quad b_i \sim N_1(0, \sigma_b^2)$$

where

$$\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T \lambda_i}}$$

Proposition 3.2.1 *given the settings in (3.2.6) then the conditional density function of $b_i|z_i, v_i$ is specified by*

$$b_i|z_i, v_i \sim N_1(\tau_i^2 1^T \Psi_i^{-1}(z_i - \Sigma_i^{1/2}\delta_i^* v_i), \tau_i^2) \tag{3.2.7}$$

where

$$\tau_i^2 = (\frac{1}{\sigma_b^2} + 1^T \Psi_i^{-1} 1)^{-1}, \quad \Psi_i = \Sigma_i^{1/2}(I - \delta_i^* \delta_i^{*t})\Sigma_i^{1/2}$$

Proof Algebraic manipulation with conditional densities yields

$$f_{b_i|z_i,v_i} = \frac{f_{z_i,b_i,v_i}}{f_{z_i,v_i}} = \frac{f_{z_i|b_i,v_i} f_{b_i} f_{v_i}}{f_{z_i|v_i} f_{v_i}} = \frac{f_{z_i|b_i,v_i} f_{b_i}}{f_{z_i|v_i}} = \frac{f_{z_i|b_i,v_i} f_{b_i}}{\int f_{z_i|b_i,v_i} f_{b_i} \mathrm{d}b_i} \tag{3.2.8}$$

now by the assumption of independence between b_i and v_i, and noting that

$$\begin{aligned} f_{z_i|v_i} =& \int f_{z_i|b_i,v_i} f_{b_i} \mathrm{d}b_i \\ =& \int N_{n_i}^{z_i}(b_i 1 + \Sigma_i^{1/2}\delta_i^* v_i, \Psi_i) N_1^{b_i}(0, \sigma_b^2) \mathrm{d}bi \quad \text{By Lemma (1.2.2)} \\ =& \int N_{n_i}^{z_i}(\Sigma_i^{1/2}\delta_i^* v_i, \Psi_i + 1\sigma_b^2 1^T) N_1^{b_i}(\tau_i^2 1^T \Psi_i^{-1}(z_i - \Sigma_i^{1/2}\delta_i^* v_i), \tau_i^2) \mathrm{d}b_i \\ =& N_{n_i}^{z_i}(\Sigma_i^{1/2}\delta_i^* v_i, \Psi_i + 1\sigma_b^2 1^T) \end{aligned} \tag{3.2.9}$$

we have by (3.2.8) and (3.2.9) and Lemma (1.2.2)

$$\begin{aligned} f_{b_i|z_i,v_i} &= \frac{N_{n_i}^{z_i}(b_i 1 + \Sigma_i^{1/2}\delta_i^* v_i, \Psi_i) N_1^{b_i}(0, \sigma_b^2)}{f_{z_i|v_i}} \\ &= \frac{N_{n_i}^{z_i}(\Sigma_i^{1/2}\delta_i^* v_i, \Psi_i + 1\sigma_b^2 1^T) N_1^{b_i}(\tau_i^2 1^T \Psi_i^{-1}(z_i - \Sigma_i^{1/2}\delta_i^* v_i), \tau_i^2)}{f_{z_i|v_i}} \\ &= \frac{f_{z_i|v_i} \times N_1^{b_i}(\tau_i^2 1^T \Psi_i^{-1}(z_i - \Sigma_i^{1/2}\delta_i^* v_i), \tau_i^2)}{f_{z_i|v_i}} \\ &= N_1^{b_i}(\tau_i^2 1^T \Psi_i^{-1}(z_i - \Sigma_i^{1/2}\delta_i^* v_i), \tau_i^2) \end{aligned} \tag{3.2.10}$$

Moreover, by applying algebraic manipulation to equation (3.2.7), and using the proof results of Proposition (1.2.3), we have

$$b_i|z_i \sim SN_1\left(\tau_i^2 1^T \Psi_i^{-1} z_i, \tau_i^2\left(1 + \tau_i^2(1^T \Psi_i^{-1}\Sigma_i^{1/2}\delta_i^*)^2\right), \lambda_{b_i}\right) \tag{3.2.11}$$

where τ_i^2 and Ψ_i as defined in (3.2.7) ,and $\lambda_{b_i} = -\tau_i 1^T \Psi_i^{-1}\Sigma_i^{1/2}\delta_i^*$.

For clarity, after the modification proposed in (3.2.3) and (3.2.5), the log-likelihood function in (2.6.3) is now defined as

$$\begin{aligned}\ell_i(\beta,\lambda_i,\Sigma_i,\sigma_{b_i}|y_i,x_i,b_i) &= \log(sn_{n_i}(z_{i1},\ldots,z_{in_i}|\Sigma_i,\lambda_i,b_i)) + \sum_{j=1}^{n_i}\log(f_{Y_{ij}}(y_{ij}|b_i,\beta,x_{ij}))\\ &- \sum_{j=1}^{n_i}\log(sn_1(z_{ij}|b_i,\lambda_{ij})) + \log(f_{b_i}(0,\sigma_b))\end{aligned} \tag{3.2.12}$$

By the characterization in (3.2.4) and (3.2.6) that are found through Proposition (1.2.3), we can introduce a dummy variable $v_i \sim HN(0,1)$ and rewrite the log-likelihood as

$$\begin{aligned}\ell_i(\beta,\lambda_i,\Sigma_i,\sigma_{b_i}|y_i,x_i,b_i) &= \log(N_{n_i}(z_{i1},\ldots,z_{in_i}|b_i 1 + \Sigma_i^{1/2}\delta_i^* v_i,\Psi_i))\\ &+ \sum_{j=1}^{n_i}\log(f_{Y_{ij}}(y_{ij}|b_i,\beta,x_{ij}))\\ &- \sum_{j=1}^{n_i}\log(N_1(z_{ij}|b_i+\delta_{ij}v_i, 1-\delta_{ij}^2)) + \log(f_{b_i}(0,\sigma_b))\\ &\propto -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i)^T\Psi_i^{-1}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i)\\ &+ \sum_{j=1}^{n_i}\log(f_{Y_{ij}}(y_{ij}|b_i,\beta,x_{ij}))\\ &- \frac{1}{2}\sum_{j=1}^{n_i}\log(1-\delta_{ij}^2) - \frac{1}{2}\sum_{j=1}^{n_i}\frac{(z_{ij}-b_i-\delta_{ij}v_i)^2}{(1-\delta_{ij}^2)}\\ &- \frac{1}{2}\log(\sigma_b) - \frac{1}{2}\frac{b_i^2}{\sigma_b^2}\end{aligned} \tag{3.2.13}$$

where $\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$, and $|\Sigma_i|$ represents the determinant. $\Psi_i = \Sigma_i^{1/2}(I - \delta_i^*\delta_i^{*t})\Sigma_i^{1/2}$ and $\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T\lambda_i}}$.

3.3 Applying the M-step

This step maximizes $Q(\theta|\theta^{(r)})$ to produce an update estimate $\theta^{(r+1)}$. Consider the score function

$$\begin{aligned}\frac{\partial}{\partial\theta}Q(\theta|\theta^{(r)}) &= \frac{\partial}{\partial\theta}\sum_{i=1}^{m}Q_i(\theta|\theta^{(r)}) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{R_i}\frac{1}{R_i}\frac{\partial}{\partial\theta}\ell_i(\theta|x_i, y_i, b_i^{(j)})\end{aligned} \tag{3.3.1}$$

Following the results obtained in equation (2.6.3), and by setting $\theta = (\beta, \sigma_b, \Sigma_i, \lambda_i)$ we have partial derivatives as:

$$\frac{\partial}{\partial\sigma_b}Q(\theta|\theta_{(r)}) = \sum_{i=1}^{m}\sum_{j=1}^{R_i}\frac{1}{R_i}\Bigg(-\frac{1}{2\sigma_b} + \frac{b_i^{2(j)}}{\sigma_b^3}\Bigg) \tag{3.3.2}$$

maximizing over the domain by setting the above equation to zero yields

$$\hat{\sigma_b^2} = \frac{2}{m}\sum_{i=1}^{m}\sum_{j=1}^{R_i}\frac{b_i^{2(j)}}{R_i} \tag{3.3.3}$$

Moreover,

$$\frac{\partial^2}{\partial\sigma_b^2}Q(\theta|\theta^{(r)}) = \sum_{i=1}^{m}\sum_{j=1}^{R_i}\frac{1}{R_i}\Bigg(\frac{1}{2\sigma_b^2} - \frac{3b_i^{2(j)}}{\sigma_b^4}\Bigg) \tag{3.3.4}$$

$$I(\sigma_b) = -E\Bigg(\frac{\partial^2}{\partial\sigma_b^2}Q(\theta|\theta^{(r)})\Bigg) = \frac{m5}{2\sigma_b^2} \tag{3.3.5}$$

where $I(.)$ is the Fisher information coefficient, and,

$$\frac{\partial^2}{\partial\sigma_b\partial\beta}Q(\theta|\theta^{(r)}) = 0 \tag{3.3.6}$$

$$\frac{\partial^2}{\partial \beta \partial \sigma_b} Q(\theta|\theta^{(r)}) = 0 \tag{3.3.7}$$

Moreover, since the term involving the parameter ϕ_i in the log-likelihood is separate from terms involving β and σ_b, we have

$$\frac{\partial^2}{\partial \phi \partial \beta} Q(\theta|.) = \frac{\partial^2}{\partial \phi \partial \sigma_b} Q(\theta|.) = \frac{\partial^2}{\partial \sigma_b \partial \phi} Q(\theta|.) = \frac{\partial^2}{\partial \beta \partial \phi} Q(\theta|.) = 0 \tag{3.3.8}$$

Form (3.3.4), (3.3.6), (3.3.7) and (3.3.8), the Hessian matrix is

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 Q(\theta|.)}{\partial \beta^2} & \frac{\partial^2 Q(\theta|.)}{\partial \beta \partial \sigma_b} & \frac{\partial^2 Q(\theta|.)}{\partial \beta \partial \phi} \\ \frac{\partial^2 Q(\theta|.)}{\partial \sigma_b \partial \beta} & \frac{\partial^2 Q(\theta|.)}{\partial \sigma_b^2} & \frac{\partial^2 Q(\theta|.)}{\partial \sigma_b \partial \phi} \\ \frac{\partial^2 Q(\theta|.)}{\partial \phi \partial \beta} & \frac{\partial^2 Q(\theta|.)}{\partial \phi \partial \sigma_b} & \frac{\partial^2 Q(\theta|.)}{\partial \phi^2} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 Q(\theta|.)}{\partial \beta^2} & 0 & 0 \\ 0 & \frac{\partial^2 Q(\theta|.)}{\partial \sigma_b^2} & 0 \\ 0 & 0 & \frac{\partial^2 Q(\theta|.)}{\partial \phi^2} \end{bmatrix} \tag{3.3.9}$$

The subsections of section (3.5) present a closed form of the parameter β maximization scheme. Moreover, since the partial derivatives of the other parameters are not quite trivial, a grid search algorithm is applied to maximize the likelihood over the parameter space.

3.4 Algorithm

The algorithm consists of the following steps

(i) Obtain an initial estimate of $\theta^{(0)}$ from the complete likelihood case and set $b_i^{(0)} = 0$.

(ii) At the $(r+1)$-th iteration, obtain MC sample on b_i.

(iii) At the $(r+1)$-th iteration, obtain the updated estimates $\theta^{(r+1)}$ using complete data optimization technique as

$$\theta^{(r+1)} = \arg\max_{\theta^{(r)}} Q(\theta|\theta^{(r)})$$

(iv) Continue steps (ii) & (iii) until convergence.

For more details see algorithm (1) in appendix (.1).

3.5 Special cases of EM algorithm

This section presents a compete E-step and M-step for the distributions considered in section (3.2) and (3.3) under different link functions $\eta(.)$ by following the GLM framework that is proposed in section (2.4). In each section an explicit marginal density function of $Y_i|b_i$ is stated along with the log-likelihood characterized in (3.2.13) and some partial derivatives. $f_{Y_{ij}}(y_{ij}|.)$ represents the marginal density of unit i, observation j of the joint density $Y_i|b_i$, $D(.)$ is a diagonal matrix and $I(.)$ is the Fisher information coefficient.

3.5.1 Exponential marginal density with link $\eta(x) = e^x$

$$f_{Y_{ij}}(y_{ij}|x_{ij}, b_i, \beta) = \exp\{\frac{-y_{ij}}{\eta(x_{ij}\beta + b_i)} - \log(\eta(x_{ij}\beta + b_i))\}, \quad y_{ij} \geq 0 \qquad (3.5.1)$$

$$\begin{aligned}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) \propto & -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i)^T \Psi_i^{-1}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i) \\ & - \frac{1}{2}\sum_{j=1}^{n_i}\log(1-\delta_{ij}^2) - \frac{1}{2}\sum_{j=1}^{n_i}\frac{(z_{ij} - b_i - \delta_{ij}v_i)^2}{(1-\delta_{ij}^2)} \\ & + \sum_{j=1}^{n_i}\{-y_{ij}e^{-x_{ij}\beta - b_i} - x_{ij}\beta - b_i\} \\ & - \frac{1}{2}\log(\sigma_b) - \frac{1}{2}\frac{b_i^2}{\sigma_b^2}\end{aligned} \tag{3.5.2}$$

where $\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$, and $|\Sigma_i|$ represents the determinant. $\Psi_i = \Sigma_i^{1/2}(I - \delta_i^*\delta_i^{*t})\Sigma_i^{1/2}$ and $\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T\lambda_i}}$

Therefore, the marginal partial derivatives defined in (3.3.1) become

$$\frac{\partial}{\partial\beta}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) = \sum_{j=1}^{n_i} x_{ij}\{y_{ij}e^{-x_{ij}\beta - b_i} - 1\} \tag{3.5.3}$$

$$\frac{\partial^2}{\partial\beta^2}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) = \sum_{j=1}^{n_i} -x_{ij}^2 y_{ij} e^{-x_{ij}\beta - b_i} \tag{3.5.4}$$

$$I(\beta) = \sum_{i=1}^{m} -E\left(\frac{\partial^2}{\partial\beta^2}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i)\right) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}^2 \tag{3.5.5}$$

$$\hat{\beta} = -(X^TX)^{-1}X^T(\log(D^{-1}(Y)1) + bI) \tag{3.5.6}$$

where $1 = (1, 1, \ldots, 1)^T$.

3.5.2 Exponential marginal density with link $\eta(x) = x^2$

$$f_{Y_{ij}}(y_{ij}|x_{ij}, b_i, \beta) = \exp\{\frac{-y_{ij}}{\eta(x_{ij}\beta + b_i)} - \log(\eta(x_{ij}\beta + b_i))\}, \quad y_{ij} \geq 0 \tag{3.5.7}$$

$$\begin{aligned}\ell_i(\beta,\lambda_i,\Sigma_i,\sigma_b|y_i,x_i,b_i) \propto & -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - b_i1 - \Sigma_i^{1/2}\delta_i^* v_i)^T\Psi_i^{-1}(z_i - b_i1 - \Sigma_i^{1/2}\delta_i^* v_i) \\ & -\frac{1}{2}\sum_{j=1}^{n_i}\log(1-\delta_{ij}^2) - \frac{1}{2}\sum_{j=1}^{n_i}\frac{(z_{ij}-b_i-\delta_{ij}v_i)^2}{(1-\delta_{ij}^2)} \\ & +\sum_{j=1}^{n_i}\{-y_{ij}(x_{ij}\beta+b_i)^{-2} - 2\log(x_{ij}\beta+b_i)\} \\ & -\frac{1}{2}\log(\sigma_b) - \frac{1}{2}\frac{b_i^2}{\sigma_b^2}\end{aligned} \tag{3.5.8}$$

where $\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$, and $|\Sigma_i|$ represents the determinant. $\Psi_i = \Sigma_i^{1/2}(I - \delta_i^*\delta_i^{*t})\Sigma_i^{1/2}$ and $\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T\lambda_i}}$

Therefore, the marginal partial derivatives defined in (3.3.1) becomes

$$\frac{\partial}{\partial\beta}\ell_i(\beta,\lambda_i,\Sigma_i,\sigma_b|y_i,x_i,b_i) = -\sum_{j=1}^{n_i}2x_{ij}\{\frac{y_{ij}}{(x_{ij}\beta+b_i)^3} - \frac{1}{x_{ij}\beta+b_i}\} \tag{3.5.9}$$

$$\frac{\partial^2}{\partial\beta^2}\ell_i(\beta,\lambda_i,\Sigma_i,\sigma_b|y_i,x_i,b_i) = -\sum_{j=1}^{n_i}2x_{ij}^2\{\frac{3y_{ij}}{(x_{ij}\beta+b_i)^4} - \frac{1}{(x_{ij}\beta+b_i)^2}\} \tag{3.5.10}$$

$$I(\beta) = \sum_{i=1}^{m} -E\left(\frac{\partial^2}{\partial\beta^2}\ell_i(\beta,\lambda_i,\Sigma_i,\sigma_b|y_i,x_i,b_i)\right) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{4x_{ij}^2}{(x_{ij}\beta+b_i)^2} \tag{3.5.11}$$

3.5.3 Exponential marginal density with link $\eta(x) = x^{-1}$

$$f_{Y_{ij}}(y_{ij}|x_{ij},b_i,\beta) = \exp\{\frac{-y_{ij}}{\eta(x_{ij}\beta+b_i)} - \log(\eta(x_{ij}\beta+b_i))\}, \quad y_{ij} \geq 0 \tag{3.5.12}$$

$$\begin{aligned}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) \propto & -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i)^T \Psi_i^{-1}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i) \\ & - \frac{1}{2}\sum_{j=1}^{n_i}\log(1-\delta_{ij}^2) - \frac{1}{2}\sum_{j=1}^{n_i}\frac{(z_{ij} - b_i - \delta_{ij}v_i)^2}{(1-\delta_{ij}^2)} \\ & + \sum_{j=1}^{n_i}\{-y_{ij}(x_{ij}\beta + b_i) + \log(x_{ij}\beta + b_i)\} \\ & - \frac{1}{2}\log(\sigma_b) - \frac{1}{2}\frac{b_i^2}{\sigma_b^2}\end{aligned} \tag{3.5.13}$$

where $\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$, and $|\Sigma_i|$ represents the determinant. $\Psi_i = \Sigma_i^{1/2}(I - \delta_i^* \delta_i^{*t})\Sigma_i^{1/2}$ and $\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T\lambda_i}}$

Therefore, the marginal partial derivatives defined in (3.3.1) become

$$\frac{\partial}{\partial\beta}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) = -\sum_{j=1}^{n_i} x_{ij}\{y_{ij} - \frac{1}{x_{ij}\beta + b_i}\} \tag{3.5.14}$$

$$\frac{\partial^2}{\partial\beta^2}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) = -\sum_{j=1}^{n_i}\frac{x_{ij}^2}{(x_{ij}\beta + b_i)^2} \tag{3.5.15}$$

$$I(\beta) = \sum_{i=1}^{m} -E\left(\frac{\partial^2}{\partial\beta^2}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i)\right) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{x_{ij}^2}{(x_{ij}\beta + b_i)^2} \tag{3.5.16}$$

$$\hat{\beta} = (X^T X)^{-1} X^T (D^{-1}(Y)1 - b1) \tag{3.5.17}$$

where $1 = (1, 1, \ldots, 1)^T$.

3.5.4 Gamma marginal density with link $\eta(x) = e^x$

Similar to the above, using equation (2.4.4), we have

$$f_{Y_{ij}}(y_{ij}|x_{ij}, b_i, \beta) = \exp\{ - ke^{(-x_{ij}\beta - b_i)}y_{ij} - k(x_{ij}\beta + b_i) + k\log(k) \\ - \log(\Gamma(k)) + (k-1)\log(y_{ij})\} \tag{3.5.18}$$

and the unit i specific marginal log-likelihood defined in (2.6.3) becomes,

$$\begin{aligned} \ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) \propto & -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i)^T \Psi_i^{-1}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i) \\ & - \frac{1}{2}\sum_{j=1}^{n_i}\log(1-\delta_{ij}^2) - \frac{1}{2}\sum_{j=1}^{n_i}\frac{(z_{ij} - b_i - \delta_{ij}v_i)^2}{(1-\delta_{ij}^2)} \\ & + \sum_{j=1}^{n_i}\{-ke^{(-x_{ij}\beta - b_i)}y_{ij} - k(x_{ij}\beta + b_i) + k\log(k) \\ & - \log(\Gamma(k)) + (k-1)\log(y_{ij})\} \\ & - \frac{1}{2}\log(\sigma_b) - \frac{1}{2}\frac{b_i^2}{\sigma_b^2} \end{aligned} \tag{3.5.19}$$

where $\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$, and $|\Sigma_i|$ represents the determinant. $\Psi_i = \Sigma_i^{1/2}(I - \delta_i^*\delta_i^{*t})\Sigma_i^{1/2}$ and $\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T\lambda_i}}$

Therefore, the marginal partial derivatives defined in (3.3.1) become

$$\frac{\partial}{\partial\beta}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) = \sum_{j=1}^{n_i} kx_{ij}\{y_{ij}e^{-x_{ij}\beta - b_i} - 1\} \tag{3.5.20}$$

$$I(\beta) = \sum_{i=1}^{m} -E\left(\frac{\partial^2}{\partial\beta^2}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i)\right) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} kx_{ij}^2 \tag{3.5.21}$$

and similar to equation (3.5.22), we have

$$\hat{\beta} = -(X^T X)^{-1} X^T (\log(D^{-1}(Y)1) + bI) \tag{3.5.22}$$

where $1 = (1, 1, \ldots, 1)^T$

3.5.5 Gamma marginal density with link $\eta(x) = x^2$

$$\begin{aligned} f_{Y_{ij}}(y_{ij}|x_{ij}, b_i, \beta) = \exp\{ &- k(x_{ij}\beta + b_i)^{-2} y_{ij} + 2k \log(x_{ij}\beta + b_i) \\ &+ k \log(k) - \log(\Gamma(k)) + (k-1)\log(y_{ij})\} \end{aligned} \tag{3.5.23}$$

and the unit i specific marginal log-likelihood defined in (2.6.3) becomes,

$$\begin{aligned} \ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) \propto &-\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i)^T \Psi_i^{-1}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i) \\ &- \frac{1}{2}\sum_{j=1}^{n_i}\log(1-\delta_{ij}^2) - \frac{1}{2}\sum_{j=1}^{n_i}\frac{(z_{ij} - b_i - \delta_{ij}v_i)^2}{(1-\delta_{ij}^2)} \\ &+ \sum_{j=1}^{n_i}\{-k(x_{ij}\beta + b_i)^{-2}y_{ij} + 2k\log(x_{ij}\beta + b_i) + k\log(k) \\ &- \log(\Gamma(k)) + (k-1)\log(y_{ij})\} \\ &- \frac{1}{2}\log(\sigma_b) - \frac{1}{2}\frac{b_i^2}{\sigma_b^2} \end{aligned} \tag{3.5.24}$$

where $\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$, and $|\Sigma_i|$ represents the determinant. $\Psi_i = \Sigma_i^{1/2}(I - \delta_i^* \delta_i^{*t})\Sigma_i^{1/2}$ and $\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T\lambda_i}}$

Therefore, the marginal partial derivatives defined in (3.3.1) become

$$\frac{\partial}{\partial\beta}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_b|y_i, x_i, b_i) = \sum_{j=1}^{n_i} 2kx_{ij}\{\frac{y_{ij}}{(x_{ij}\beta + b_i)^3} + \frac{1}{x_{ij}\beta + b_i}\} \tag{3.5.25}$$

$$I(\beta) = \sum_{i=1}^{m} -E\left(\frac{\partial^2}{\partial\beta^2}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_b|y_i, x_i, b_i)\right) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{4kx_{ij}^2}{(x_{ij}\beta + b_i)^2} \tag{3.5.26}$$

3.5.6 Gamma marginal density with link $\eta(x) = x^{-1}$

Similar to the above, using equation (2.4.4), we have

$$f_{Y_{ij}}(y_{ij}|x_{ij}, b_i, \beta) = \exp\{-k(x_{ij}\beta + b_i)y_{ij} + k\log(x_{ij}\beta + b_i) + k\log(k) - \log(\Gamma(k)) + (k-1)\log(y_{ij})\} \tag{3.5.27}$$

and the unit i specific marginal log-likelihood defined in (2.6.3) becomes,

$$\begin{aligned}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) \propto & -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i)^T \Psi_i^{-1}(z_i - b_i 1 - \Sigma_i^{1/2}\delta_i^* v_i) \\ & - \frac{1}{2}\sum_{j=1}^{n_i}\log(1-\delta_{ij}^2) - \frac{1}{2}\sum_{j=1}^{n_i}\frac{(z_{ij} - b_i - \delta_{ij}v_i)^2}{(1-\delta_{ij}^2)} \\ & + \sum_{j=1}^{n_i}\{-k(x_{ij}\beta + b_i)y_{ij} + k\log(x_{ij}\beta + b_i) + \\ & k\log(k) - \log(\Gamma(k)) + (k-1)\log(y_{ij})\} \\ & - \frac{1}{2}\log(\sigma_b) - \frac{1}{2}\frac{b_i^2}{\sigma_b^2}\end{aligned} \tag{3.5.28}$$

where $\delta_{ij} = \frac{\lambda_{ij}}{\sqrt{1+\lambda_{ij}^2}}$, and $|\Sigma_i|$ represents the determinant. $\Psi_i = \Sigma_i^{1/2}(I - \delta_i^*\delta_i^{*t})\Sigma_i^{1/2}$ and $\Sigma_i^{1/2}\delta_i^* = \frac{\lambda_i}{\sqrt{1+\lambda_i^T\lambda_i}}$

Therefore, the marginal partial derivatives defined in (3.3.1) become

$$\frac{\partial}{\partial\beta}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i) = \sum_{j=1}^{n_i} kx_{ij}\{-y_{ij} + \frac{1}{x_{ij}\beta + b_i}\} \tag{3.5.29}$$

$$I(\beta) = \sum_{i=1}^{m} -E\left(\frac{\partial^2}{\partial\beta^2}\ell_i(\beta, \lambda_i, \Sigma_i, \sigma_{b_i}|y_i, x_i, b_i)\right) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\frac{kx_{ij}^2}{(x_{ij}\beta + b_i)^2} \tag{3.5.30}$$

and similarly

$$\hat{\beta} = (X^T X)^{-1} X^T (D^{-1}(Y)1 - b1) \tag{3.5.31}$$

where $1 = (1, 1, \ldots, 1)^T$

Chapter 4

Simulation and application

4.1 Simulation design

To assess the efficiency of the proposed likelihood and model above multiple key statistics are needed. Two different model setting are chosen for inference and assessment, a univariate model and a bivariate model. A unified simulation structure is selected as the number of units to be fixed to 5, such that $i = 1, \ldots, 5$, and under each simulation the number of observation is chosen randomly from uniform distribution $n_i \sim \mathcal{U}(50, 250)$. Once the number of units and observations per unit are decided, the per unit variance-covariance matrix is constructed as a first order autoregressive model as in (2.3.6), with a choice of $\phi_i = 0.15 \quad \forall i$. Note that in the case of $\phi_i = 0$ the autoregressive structure reduces to an independent multivariate distribution. Finally, $\sigma_b = 2$.

The following two subsections discuss the model specific settings.

4.1.1 Univariate model

X_i is generated from $N_{n_i}(0, I_{n_i \times n_i})$, $\phi_i = 0.15 \quad \forall i$, $\beta = 3$, and $b_i \sim N_1(0, \sigma_b = 2)$. Moreover, the time difference per observation within each unit is set to a unit difference, i.e, in (2.3.6), $e^{-\phi_i |t_{ij} - t_{ik}|}$ would reduce to

$$e^{-\phi_i |t_{ij} - t_{ik}|} = \begin{cases} e^{-\phi_i} & \text{if } |j-k| = 1 \\ 1 & \text{if } j = k \end{cases} \tag{4.1.1}$$

Therefore the response variable is generated from

$$Y_i \sim MV(\eta(X_i \beta + b_i), \Sigma_i(\phi_i)) \tag{4.1.2}$$

with a chosen multivariate distribution and link function as in section (3.5).

4.1.2 Bivariate model

This model investigates the convergence under an extra binary variable t_{ij}, which in some cases could represent a measurement deviation under th existence of certain events. Here $t_{ij} = 1$ for $j \leq 150$ and $t_{ij} = 0$ for all $j > 150$. Similarly as above X_i is generated from $N_{n_i}(0, I_{n_i \times n_i})$, $\phi_i = 0.15 \quad \forall i$, $\beta = (3, 2)^T$, and $b_i \sim N_1(0, \sigma_b = 2)$. Moreover, the time difference per observation within each unit is set to a unit difference. The response variable is then is generated from

$$Y_i \sim MV(\eta([X_i, t_i]\beta + b_i), \Sigma_i(\phi_i)) = MV(\eta(3X_i + 2t_i + b_i), \Sigma_i(\phi_i)) \tag{4.1.3}$$

The original correlation matrix Σ_i is used alongside original values of λ_i and b_i

to simulate a set of random variables $Z_i|b_i$ defined in (3.2.5). Using those random variables and the inverse transformation method introduced in (2.2.2) along with the desired marginal densities and link function to generate $Y_i|b_i$. Finally to initialize each simulation we set the initial vectors to $\beta^{(0)} = 0$, $\lambda_i^{(0)} = 0$, $\sigma^2_{b^{(0)}} = 1$ and $\phi_i^{(0)} = 0.5$. In each iteration, the Monte Carlo sampling from $b^{(k)}|Z$ is set to 300 and gradually increases while algorithm (1) is run until convergence.

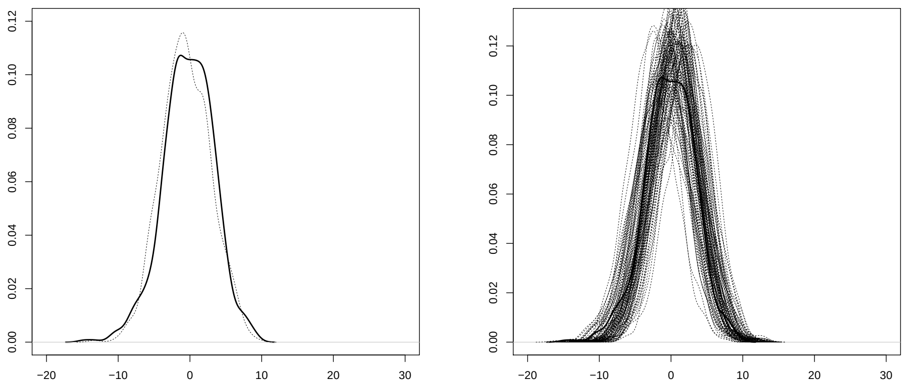
4.2 Simulation under special cases of link and distribution function

Similar to section (3.5) and the special cases presented, this simulation presents both models, a univariate and a bivariate, under an exponential link function $\eta(x) = e^x$.

Note that all tables in this section represent a simulation of 100 Monte Carlo data sets, where MC Mean and MC SD represent the Monte Carlo mean and standard deviation. MSE represents the average standard error between Monte Carlo simulation and the true value of the parameter. EC represents the empirical coverage probability computed using Fisher information matrix assuming a 95% confidence interval. True values are shown in parentheses (.) next to the parameter symbol, i.e, $\beta(3)$ implies original β parameter is set to 3.

4.2.1 Exponential marginal density

Before jumping to a complete simulation analysis, figure (4.1) below depict the convergence approximation graphically of the proposed model under a single simulated data set, using a univariate model with an exponential link and distribution function. In addition, (4.2) depicts similar analysis under a multivariate model.



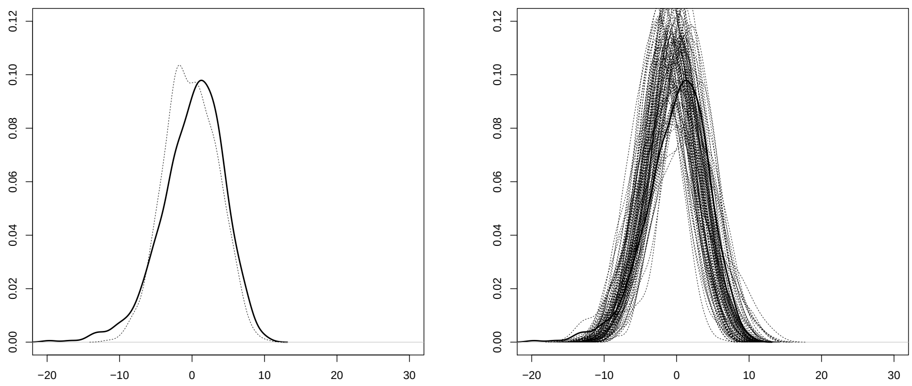
(a) A single replication (b) 100 MC replications

Figure 4.1: A univariate model with an exponential link and distribution function, where the log density of Y versus $\log(\eta(.))$ are plotted on the x-axis; in bold and dotted lines respectively. (4.1a) compares a single replication of the estimated model versus the real model, while (4.1b) is a 100 Monte Carlo replications versus the real model.

Moreover, table (4.1) and (4.2) represent the parameter estimation under the univariate and multivariate model respectively.

Table 4.1: Univariate model under and exponential distribution such that $E[Y_i] = \mu_i = e^{X_i\beta+b_i}$ and a variance-covariance matrix $\Sigma_i(\phi)$ with $\phi_i = 0.15$, $\forall i$

Parameters	MC Mean	MC SD	MSE	EC
$\beta(3)$	2.91	0.02	0.01	0.54
$E[b](0)$	0.08	0.18	-	-
$\sigma_b(2)$	2.15	0.41	0.19	0.99
$\phi(0.15)$	0.45	0.23	0.15	0



(a) A single replication (b) 100 MC replications

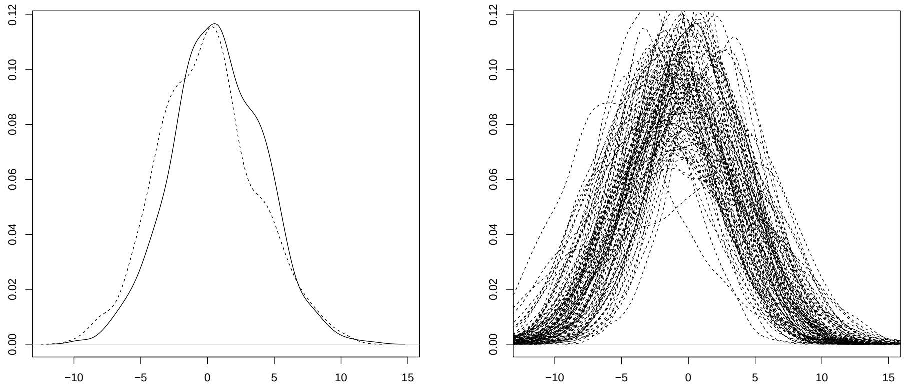
Figure 4.2: A multivariate model with an exponential link and distribution function, where the log density of Y versus $\log(\eta(.))$ are plotted on the x-axis; in bold and dotted lines respectively. (4.2a) compares a single replication of the estimated model versus the real model, while (4.2b) is a 100 Monte Carlo replications versus the real model.

Table 4.2: Multivariate model under and exponential distribution such that $E[Y_i] = \mu_i = e^{X_i\beta+b_i}$ and a variance-covariance matrix $\Sigma_i(\phi)$ with $\phi_i = 0.15$, $\forall i$

Parameters	MC Mean	MC SD	MSE	EC
$\beta_1(3)$	3.03	0.16	0.002	0.99
$\beta_2(2)$	1.53	0.64	0.63	0.20
$E[b](0)$	-0.04	0.09	-	-
$\sigma_b(2)$	1.99	0.43	0.18	0.93
$\phi(0.15)$	0.38	0.19	0.09	-

4.2.2 Gamma marginal density

Similar to the earlier subsection, this section offers a graphical example of a single simulation under an exponential link function and a gamma in figure (4.3) distribution with shape parameter $k = 3$. To be concise only simulations under a multivariate model are shown.



(a) A single replication (b) 100 MC replications

Figure 4.3: A multivariate model with an exponential link and gamma distribution function, where the log density of Y versus $\log(\eta(.))$ are plotted on the x-axis; in bold and dotted lines respectively. (4.3a) compares a single replication of the estimated model versus the real model, while (4.3b) is a 100 Monte Carlo replications versus the real model. The shape parameter $k = 3$.

Table 4.3: Multivariate model under a gamma distribution such that $E[Y_i] = \mu_i = e^{X_i\beta+b_i}$, where k is the shape parameter and is fixed to 3. A variance-covariance matrix $\Sigma_i(\phi)$ with $\phi_i = 0.15$, $\forall i$

Parameters	MC Mean	MC SD	MSE	EC
$\beta_1(3)$	2.97	0.06	0.003	0.92
$\beta_2(2)$	1.15	1.14	2.01	0.1
$E[b](0)$	0.11	0.11	-	-
$\sigma_b(2)$	2.37	0.45	0.42	0.96
$\phi(0.15)$	0.33	0.27	0.11	-

4.3 An application

For comparison reasons we consider the same data set that Zhang and Davidian [2001] and Arellano-Valle et al. [2005] have both used, in particular, the Framingham Heart Study, which consists of longitudinal data for a wide set of cohorts. Zhang and Davidian [2001] used a linear mixed model approach to study the change of cholesterol levels over time within patients. The set includes 200 randomly selected participants along with their gender, age and cholesterol levels, where the cholesterol levels are measured at the beginning of the study and every two years for the total of 10 years. The model they used is

$$Y_{ij} = \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij} \tag{4.3.1}$$

Here Y_{ij} is the cholesterol level divided by 100 at the j^{th} time for unit i and t_{ij} is $\frac{\text{time}-5}{10}$, with time measured in years from baseline. $\epsilon_{ij} \overset{\text{iid}}{\sim} N_1(0, \sigma^2)$; age_i is age at baseline; sex_i is a gender indicator$(0 = \text{female}, 1 = \text{male})$. $\beta = (\beta_1, \beta_2)$ are the fixed effects coefficients, and $b_i = (b_{0i}, b_{1i})$ are the unit specific random effects coefficient.

Since the modeling approach proposed in chapter (2), takes into account the t_{ij} variable as a part of the variance-covariance matrix Σ_i then using similar variables proposed in previous paragraph the modified model would be

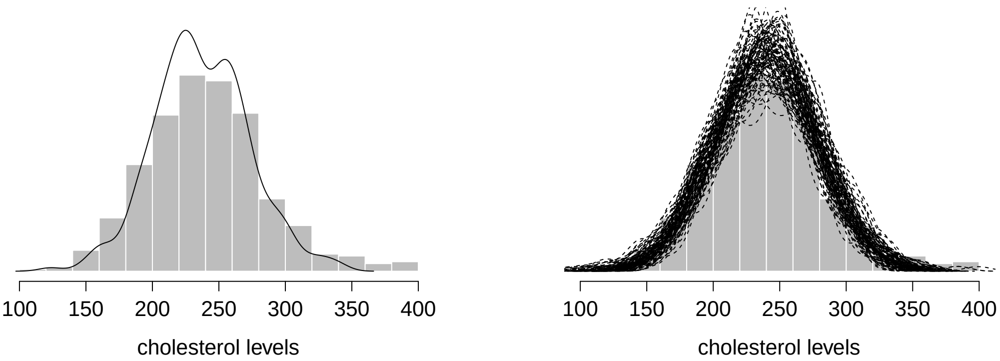
$$Y_i \sim MV(\beta_1 age_i + \beta_2 sex_i + b_i, \Sigma_i(\phi_i, t_{ij})) \tag{4.3.2}$$

b_i here is the unit specific random effect proposed first in equation (3.2.4); t_{ij} is as in previous paragraph $\frac{\text{time}-5}{10}$, and the correlation coefficients are defined as

$$Corr(Y_{ij}, Y_{ik}) = e^{-\phi_i t_{ij}} \tag{4.3.3}$$

modeling is performed with a gamma distribution and an exponential link function as in subsection (3.5.4).

Figure (4.4a)represents a histogram of cholesterol levels of the 200 randomly selected patients with the solid line as the fitted model under the proposed settings. Moreover, figure (4.4b) shows the same histogram versus a 100 MC replications of b_i.



(a) A single replication (b) 100 MC replications

Figure 4.4: Fitting of Framingham Heart Study cholesterol data with model (4.3.2) using an exponential link and gamma distribution function, the shape parameter $k = 3$. The solid lines are the fitted model, while the histogram shows the frequency distribution of cholesterol levels.

Similar to the simulation section above, table (4.4) presents the parameter estimates and standard errors which are calculated as $SE(\theta_{\text{MLE}}) = \frac{1}{\sqrt{I(\theta_{\text{MLE}})}}$, where I is the Fisher Information coefficient of the maximum likelihood estimate of parameter θ.

Table 4.4: Fitting of Framingham Heart Study cholesterol data with model (4.3.2) using an exponential link and gamma distribution function, the shape parameter $k = 3$.

Parameters	Estimate	SE
β_1	0.002	0.0005
β_2	0.22	0.02
σ_b	0.996	0.045
ϕ	0.05	-
-log-likelihood	-1966.86	
AIC	-7.93657	
BIC	5.2567	

Intuitively, one seeks comparison results with different models. As mentioned earlier, Arellano-Valle et al. [2005] has fitted the Framingham Heart Study cholesterol data under a mixture of Gaussian and skew-normal distribution for the random effects and residuals. The additive structure of the model is shown in equation (4.3.1) with a bivariate random effect, while the presented model in (4.3.2) uses one. For this reason, table (4.5) presents a numerical comparison of average mean square error (Ave MSE) of two of Arellano-Valle et al. [2005]'s models to the presented copula-driven model, where the average is taken over 10000 runs.

A.V Model 1 A model with independent multivariate normal distribution for errors and multivariate skew-normal distribution for random effects with $\lambda_b = (\lambda_{b_1}, \lambda_{b_2})^T$.

A.V Model 2 A model with independent multivariate skew-normal distribution for errors with common shape parameter between groups and multivariate symmetric normal distribution for the random effects.

The estimated coefficients are presented in Table 3 in Arellano-Valle et al. [2005].

It is clear that the average MSE of Arellano-Valle et al. [2005] surpasses the fit of the proposed model; keep in mind, that Arellano-Valle et al. [2005] model have an extra random effect and parameters to estimate. Nevertheless, this is the first step to estimate mixed models via a skew-normal copula, and future research is inevitable for better fits, and most importantly, for the integration of a random effects design matrix.

Table 4.5: Fitting of Framingham Heart Study cholesterol data comparison table with Arellano-Valle et al. [2005]

Factor	A.V Model 1	A.V Model 2	Copula Driven-Model
Ave MSE	0.3415	0.3539	3.691

Chapter 5

Conclusion and final remarks

After characterizing the univariate and multivariate skew-normal distributions in Chapter (1), we were able to use such findings to construct a skew-normal copula. Chapter (2) works out all the details needed to simulate a general multivariate distribution. Later chapters apply these findings to the exponential family distribution, and present a complete derivation of the likelihood for a gamma and an exponential distribution under different link functions. The numerical approximations and results in chapter (4.1) confirm the premise of this paper.

In future research, one can investigate different prior distribution functions for the b_i parameter in equation (3.2.4). For example, considering a skew-normal distribution for the random effects. Another possibility is to consider different distributions for b_i between units in longitudinal data, the possibility of such structure comes from section (2.6) where the additive characteristic of the likelihood permits such dynamics between units of data.

Appendix A

.1 MCEM algorithm diagram

A more detailed step-by-step walk through to the algorithm presented in section (3.4).

Algorithm 1 MCEM algorithm

Initialize $\theta^{(0)} = (\beta^{(0)} = 0, \Sigma_i^{(0)} = I_{(n_i \times n_i)}, \sigma_{b_i}^{(0)} = 1, \lambda_i^{(0)} = 0)$
Run steps (i)-(iv)

(i) Set all $R = \{R_1, R_2, \ldots, R_m\}$ to desirable large numbers.

(ii) Generate $v_i^{(0)}$ and $Z_i^{(0)} | b_i^{(0)}, v_i^{(0)}, \lambda_i^{(0)}$ from (3.2.6).

(iii) Generate an R_i Monte Carlo sample for $b_i^{(0)} | Z_i^{(0)}, v_i^{(0)}$ as in (3.2.7).

(iv) Compute $Q_i(\theta|\theta^{(0)})$ as in (3.2.1), and $Q(\theta|\theta^{(0)})$ as in (3.2.2)

For the selected distribution function $F(.)$ and the link function $\eta(.)$ run the following convergence loop in next algorithm number (2).

Algorithm 2 MCEM convergence loop

while $Q(\theta|\theta^{(r+1)}) \geq Q(\theta|\theta^{(r)})$, for $r \geq 0$ **do**

(1) - Compute the M-step as in section (3.3) as follows

(a) Compute $\hat{\beta}^{(r+1)}$given $\theta^{(r)}$.

(b) $\theta^{(r+1)} = \arg\max_{\theta^{(r)}} Q(\theta|\theta^{(r)}, \hat{\beta}^{(r+1)})$.

(2) - Compute the E-step $Q(\theta|\theta^{(r+1)})$ as in (3.2.2) given the new $\theta^{(r+1)}$.

(3) - Check the while condition is still valid. If valid continue, otherwise reset $\hat{\beta}^{(r+1)} = \hat{\beta}^{(r)}$ and $\hat{\theta}^{(r+1)} = \hat{\theta}^{(r)}$.

(4) - Generate $Z_i^{(r+1)}|b_i^{(r)}, v_i$ as in (3.2.6).

(5) - Generate a new Monte Carlo sample $b_i^{(r+1)}|Z_i^{(r+1)}, v_i$ as in (3.2.7).

(6) - Repeat steps (1) - (6).

end while

References

R. Arellano-Valle and M. G. Genton. Fundamental skew distributions. *Journal of Multivariate Analysis*, 96:93–116, 2005. 8, 9

R. Arellano-Valle, G. Del Pino, and E. San Martin. Definition and probabilistic properties of skew distribution. *Statistic and Probability Letters*, 58:111–121, 2002. 7

R. Arellano-Valle, H. Bolfarine, and V. Lachos. Skew-normal linear mixed models. *Journal of Data Science*, 3:415–438, 2005. iii, ix, 3, 9, 10, 11, 16, 29, 49, 52, 53

A. Azzalini. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12:171–178, 1985. 3, 5, 6, 30

A. Azzalini and A. Capitanio. Statistical application of the multivariate skew normal distribution. *Journal of the Royal Statistical Society*, 61:579–602, 1999. 6

A. Azzalini and A. Dalle-Valle. The multivariate skew-normal distribution. *Biometrika*, 83:715–726, 1996. 6, 9

S. M. Butler and T. A. Louis. Random effects models with non-parametric priors. *Statistics in Medicine*, 11:1981–2000, 1992. 2

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39 (1):1–38, 1977. 28, 29

Y. Dodge, editor. *The Oxford Dictionary of Statistical Terms*. Oxford University Press, 2003. 13

RA Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918. 1

M. G. Gallant and D. W. Nychka. Semi nonparametric maximum likelihood estimation. *Econometrica*, 54:363–390, 1987. 3

L. S. Magder and S. L. Zeger. A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association*, 91:1141–1152, 1996. 3

X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:67–278, 1993. 28

Roger B. Nelsen, editor. *An Introduction to Copulas*. Springer, 1999. 13

M. A. Newton and Y. Zhang. A recursive algorithm for non-parametric analysis with missing data. 1999. 3

Joel Owen and Ramon Rabinovitch. On the class of elliptical distributions and their applications to the theory of portfolio choice. *The Journal of Finance*, 38 (3):745–752, 1983. 7

G.K. Robinson. That BLUP is a good thing: The estimation of random effects. *Statistical Science*, 6:15–32, 1991. 1

H. Tao, M. Aptal, Yandell B. S., and Newton M. A. An estimation method for the semi-parametric mixed effects model. *Biometrics*, 55:102–110, 1999. 3

G. Verbeke and E. Lesaffre. A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91:217–221, 1996. 3

G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor mans data augmentation algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990. 28

C.F Jeff Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103, 1983. 29

Daowen Zhang and Marie Davidian. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*, 57:795–802, 2001. 3, 49, 50