Jiaqi's Thesis Progress Report (Updated Apr. 17)

Jiaqi Bi[a]

[a] *Western University,*
Schulich School of Medicine & Dentistry,
Department of Epidemiology and Biostatistics

1. To Do List

1. Gibb's sampler

2. Notations

List of Notations

i	Individual index
j	Family (Cluster) index
p	Proband index
d_j	Number of events in family j
t	Some time
a	Some Time for the proband
T	Event Time
δ_{ij}	Event indicator for individual i in family j
w	The observed survival data (t, δ)
n	Number of individuals
J	Number of Families (Clusters)
m	Index of the sampled completed dataset in the MCEM
M	Number of the sampled completed dataset in the MCEM
z	Frailty term
q	q-th element of Gauss Hermite Quadrature
ω	q-th weight of Gauss Hermite Quadrature
y_q	q-th node of Gauss Hermite Quadrature
N_q	Total number of quadratures
$h(\cdot)$	Hazard fucntion
$h_0(\cdot)$	Baseline hazard function
$H(\cdot)$	Cumulative hazard fucntion
$S(\cdot)$	Survival fucntion
$A_j(\cdot)$	Ascertainment of family j into the study

Email address: `jbi23@uwo.ca` (Jiaqi Bi)

Preprint submitted to Dr. Choi & Dr. Espin-Garcia *Mar. 29, 2024*

$L(\cdot)$	Likelihood function
$\ell(\cdot)$	Log-likelihood function
$\mathscr{L}(\cdot)$	Laplace transform
$\mathbf{x}$	Covariates
$\boldsymbol{\beta}$	Model coefficients vector
$\boldsymbol{\theta}$	Parameter vector
Λ	The combination of $(\boldsymbol{\beta}, \lambda, \alpha)$
λ	Weibull shape parameter
α	Weibull scale parameter
υ	General form of the parameter in an undefined frailty distribution
k	Gamma shape and rate parameters
σ^2	Log-Normal variance parameter
ψ	Missing data distribution parameters

3. Weibull Parametric Approach

For the model efficiency of the analyses in a genetic research, a parametric survival analysis is usually chosen over semi-parametric survival analysis [1, 2]. From the beginning of the discussion, I have obtained the model, i.e., the hazard function is

$$h_{ij}(t_{ij}|\mathbf{x}_{ij}, z_j) = h_0(t_{ij}) \exp(\beta_1 x_{1,ij} + \beta_2 x_{2,ij}) z_j \tag{1}$$

There are total n_j individuals in family j, where $i = 1, ..., n_j$, and total J families that $j = 1, ..., J$. $x_{1,ij}$ is the genotype, or say mutation gene status for individual i in family j. $x_{2,ij}$ is the PRS for individual i in family j. The frailty term z_j, has a pdf of $f(z)$, which can be Gamma, log-normal, or other frailty distributions. The support of $f(z)$ is always non-negative. The Weibull baseline hazard function is defined as

$$h_0(t_{ij}) = \alpha^\lambda \lambda t_{ij}^{\lambda-1} \tag{2}$$

where λ is the shape parameter and α is the rate parameter. Let $\xi_{ij} = \exp(\beta_1 x_{1,ij} + \beta_2 x_{2,ij})$, the hazard function is

$$h_{ij}(t_{ij}|\mathbf{x}_{ij}, z_j) = \alpha^\lambda \lambda t_{ij}^{\lambda-1} \xi_{ij} z_j \tag{3}$$

The survival function $S(t)$ can be obtained through cumulative hazard function $H(t)$

$$\begin{aligned} H(t_{ij}|\mathbf{x}_{ij}, z_j) &= \int_0^t h_{ij}(u|\mathbf{x}_{ij}, z_j) du & (4) \\ &= \alpha^\lambda \xi_{ij} z_j \lambda \int_0^t u^{\lambda-1} du & (5) \\ &= \alpha^\lambda \xi_{ij} z_j \lambda \cdot \frac{1}{\lambda} t_{ij}^\lambda = \alpha^\lambda \xi_{ij} z_j t_{ij}^\lambda & (6) \end{aligned}$$

and the survival function

$$S(t_{ij}|\mathbf{x}_{ij}, z_j) = \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) = \exp(-\alpha^{\lambda}\xi_{ij}z_j t_{ij}^{\lambda}) \tag{7}$$

Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \alpha, \lambda, \upsilon\}$, where υ is the parameter for the frailty distribution of the choice. In our example dataset, $\boldsymbol{\beta} = (\beta_1, \beta_2)$. Therefore, the likelihood assuming missing data and frailties are observed can be written as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{J}\prod_{i=1}^{n_j}(\alpha^{\lambda}\lambda t_{ij}^{\lambda-1}\xi_{ij}z_j)^{\delta_{ij}}\exp(-\alpha^{\lambda}\xi_{ij}z_j t_{ij}^{\lambda}) && (8)\\ &= \prod_{j=1}^{J}\prod_{i=1}^{n_j} h(t_{ij}|\mathbf{x}_{ij}, z_j)^{\delta_{ij}}\exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) && (9)\end{aligned}$$

When there is no missing data but frailties are present, the frailty term can be integrated where the likelihood is taken to be the expectation with respect to the frailty z_j. The likelihood can be written as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{z_j}(\alpha^{\lambda}\lambda t_{ij}^{\lambda-1}\xi_{ij}z_j)^{\delta_{ij}}\exp(-\alpha^{\lambda}\xi_{ij}z_j t_{ij}^{\lambda})f(z_j)dz_j && (10)\\ &= \prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{z_j} h(t_{ij}|\mathbf{x}_{ij}, z_j)^{\delta_{ij}}\exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j))f(z_j)dz_j && (11)\end{aligned}$$

But when the missing data and the frailty both exist in the model, we will need to account for their joint distribution within the likelihood according to Herring et al. [3].

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{z_j,\mathbf{x}_{mis,ij}}(\alpha^{\lambda}\lambda t_{ij}^{\lambda-1}\xi_{ij}z_j)^{\delta_{ij}}\exp(-\alpha^{\lambda}\xi_{ij}z_j t_{ij}^{\lambda})f(z_j, \mathbf{x}_{mis,ij})dz_j d\mathbf{x}_{mis,ij} && (12)\\ &= \prod_{j=1}^{J}\prod_{i=1}^{n_j}\int_{z_j,\mathbf{x}_{mis,ij}} h(t_{ij}|\mathbf{x}_{ij}, z_j)^{\delta_{ij}}\exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j))f(z_j, \mathbf{x}_{mis,ij})dz_j d\mathbf{x}_{mis,ij} && (13)\end{aligned}$$

The following section 5 and section 6 discuss how to handle the frailty within the likelihood when there are no missing data, which are corresponding to the Equation 10 and 11. The section 7 will discuss how to handle the frailty and the missing data jointly, which is corresponding to the likelihood equation and 13.

4. Ascertainment Correction

Within a genetic study, those families are typically selected when there is an affected person called a proband. This will yield a selection bias because this is no long a case-control study, and can potentially defect the statistical power [4, 5]. It is crucial to address the ascertainment bias. Consider A as the event of being ascertained, D as the data, we then have $P(D, A|\boldsymbol{\theta}) = P(A|D, \boldsymbol{\theta})P(D|\boldsymbol{\theta})$. Also, we know A is included in D, from Baye's rule

$$P(D|\boldsymbol{\theta}) = \frac{P(D, A|\boldsymbol{\theta})}{P(A|D, \boldsymbol{\theta})} \propto \frac{L(\boldsymbol{\theta}|D)}{P(A|D, \boldsymbol{\theta})} \tag{14}$$

For each family j, the ascertainment A_j is defined to be the probability of the proband p being ascertained by the age a_{p_j} at examination, i.e., $A_j = P(T_{p_j} < a_{p_j})$ where a_{p_j} is proband's age at study entry. Applying the ascertainment correction for the log-likelihood in family j:

$$\tilde{\ell}_j(\boldsymbol{\theta}) = \ell_j(\boldsymbol{\theta}) - \log A_j(\boldsymbol{\theta}) \tag{15}$$

where $\tilde{\ell}$ is the log-likelihood with ascertainment correction, and ℓ is the crude log-likelihood. Define $\mathbf{x}_{p_j}$ the covariates for proband in family j, so we can further write the formula for the ascertainment correction within different frailty models.

5. Gamma Frailty

We can obtain the likelihood for Gamma frailty model following the instruction by Balan and Putter [6]. The Laplace transform of the frailty $z \sim \text{Gamma}(k, k)$, for the simplicity of the mathematical expression, the following Laplace transform will ignore the subscript, denote $\mathscr{L}(f(z)) = \phi(s)$ where $s = \sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij})$:

$$\begin{aligned} \phi(s) &= \int_0^\infty e^{-sz} f(z) dz & (16) \\ &= \int_0^\infty e^{-sz} \frac{k^k}{\Gamma(v)} z^{k-1} e^{-kz} dz & (17) \end{aligned}$$

Using the Gamma property: $\int_0^\infty z^{n-1} e^{-az} dz = \frac{\Gamma(n)}{a^n}$, $\phi(s)$ can be further written as

$$\phi(s) = \frac{k^k}{\Gamma(k)} \int_0^\infty e^{-(s+k)z} z^{k-1} dz = \frac{k^k}{\Gamma(k)} \cdot \frac{\Gamma(k)}{(s+k)^k} = (1 + \frac{s}{k})^{-k} \tag{18}$$

The second derivative is $\frac{d^2\phi(s)}{ds^2} = \int_0^\infty (-z)^2 e^{-sz} f(z) dz$.
The third derivative is $\frac{d^3\phi(s)}{ds^3} = \int_0^\infty (-z)^3 e^{-sz} f(z) dz$, ... Therefore, its d-th derivative, denote $\phi(s)^{(d)}$:

$$\begin{aligned} \phi(s)^{(d)} &= (-1)^d \int_0^\infty z^d e^{-sz} f(z) dz & (19) \\ &= (-1)^d \frac{(k+d-1)!}{(k-1)!(s+k)^d} (1 + \frac{s}{k})^{-k} & (20) \end{aligned}$$

Let $\boldsymbol{\theta} = (\beta_1, \beta_2, \alpha, \lambda, k)$ for Gamma frailty model, the log-likelihood is then written as

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{k} \log \left[\int_0^\infty \prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}, z_j))^{\delta_{ij}} \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) f(z_j) dz_j \right] \tag{21}$$

$$= \sum_{j=1}^{J} \log \left[\int_0^\infty \prod_{i=1}^{n_j} (z_j h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \exp(-z_j H(t_{ij}|\mathbf{x}_{ij})) f(z_j) dz_j \right] \tag{22}$$

$$= \sum_{j=1}^{J} \log \left[\prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \int_0^\infty z_j^{d_j} \exp(-z_j \sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij})) f(z_j) dz_j \right] \tag{23}$$

$$= \sum_{j=1}^{J} \log \left[\prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \frac{(k+d_j-1)!}{(k-1)!(\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) + k)^{d_j}} \Big(1 + \frac{\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij})}{k}\Big)^{-k} \right] \tag{24}$$

$$= \sum_{j=1}^{J} \log \left[\prod_{i=1}^{n_j} ((h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}}) \frac{(k+d_j-1)!}{k!k^{d_j-1}} (1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}))}{k})^{-k-d_j} \right] \tag{25}$$

$$= \sum_{j=1}^{J} \log \left[h(t_{ij}|\mathbf{x}_{ij})^{\delta_{ij}} \frac{(k+d_j-1)!}{k!k^{d_j-1}} (1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}))}{k})^{-k-d_j} \right] \tag{26}$$

$$= \sum_{j=1}^{J} \left[\sum_{i=1}^{n_j} (\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij})) + \log \Big(\frac{(k+d_j-1)!}{k!k^{d_j-1}} (1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}))}{k})^{-k-d_j} \Big) \right] \tag{27}$$

Not all probands in their study entry ages are affected, so it is crucial to apply a ascertainment correction accordingly. Denote $I(T_{p_j} < a_{p_j})$ as an indicator of the proband was affected before their entry to the study. Note we can still apply Laplace transform for the ascertainment correction, such that

$$A_j(\boldsymbol{\theta}) = \Big[1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j})\Big]^{I(T_{p_j} < a_{p_j})} S_{p_j} \Big[(a_{p_j}|\mathbf{x}_{p_j})\Big]^{1 - I(T_{p_j} < a_{p_j})} \tag{28}$$

$$= \Big[1 - \int_0^\infty S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}, z_j) f(z_j) dz_j\Big]^{I(T_{p_j} < a_{p_j})} \Big[\int_0^\infty S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}, z_j) f(z_j) dz_j\Big]^{1 - I(T_{p_j} < a_{p_j})} \tag{29}$$

$$= \Big[1 - \int_0^\infty \exp(-z_j \cdot H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})) f(z_j) dz_j\Big]^{I(T_{p_j} < a_{p_j})} \tag{30}$$

$$\times \Big[\int_0^\infty \exp(-z_j \cdot H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})) f(z_j) dz_j\Big]^{1 - I(T_{p_j} < a_{p_j})} \tag{31}$$

$$= \Big[1 - (1 + \frac{H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})}{k})^{-k}\Big]^{I(T_{p_j} < a_{p_j})} \Big[(1 + \frac{H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})}{k})^{-k}\Big]^{1 - I(T_{p_j} < a_{p_j})} \tag{32}$$

6. Log-Normal Frailty

The log-normal frailty is not the power-variance-function (PVF) family, so there is no closed form for Laplace transform or expressions for survivors. But we are able to estimate the Laplace transform using Gauss Hermite Quadrature. We typically standardize the log-normal frailty Z as

$$E(\log Z) = 0 \tag{33}$$
$$\text{Var}(\log Z) = \sigma^2 \tag{34}$$

That is, $z \sim$ log-Normal$(0, \sigma^2)$. The probability density function $f(z)$ is then

$$f(z) = \frac{1}{\sqrt{2\pi}\sigma} z^{-1} \exp(-\frac{\log(z)^2}{2\sigma^2}) \tag{35}$$

The Laplace transform is then

$$\phi(s) = \mathscr{L}(f_Z)(s) = \int_0^\infty \exp(-sz) \cdot f(z) dz \tag{36}$$

Using variable transformation, let $y = \frac{\log(z)}{\sqrt{2}\sigma}$, then $z = \exp(\sqrt{2}\sigma y)$, and $dz = \sqrt{2}\sigma \exp(\sqrt{2}\sigma y) dy$. Therefore, for d-th derivative:

$$\phi(s)^d = \int_{-\infty}^\infty z^d \exp(-sz) \cdot \frac{1}{\exp(\sqrt{2}\sigma y)\sigma\sqrt{2\pi}} \cdot \exp(-y^2) \cdot \sqrt{2}\sigma \exp(\sqrt{2}\sigma y) dy \tag{37}$$
$$= \int_{-\infty}^\infty \exp(\sqrt{2}\sigma y)^d \exp(-s \exp(\sqrt{2}\sigma y)) \cdot \frac{1}{\sqrt{\pi}} \exp(-y^2) dy \tag{38}$$

Definition 1 (Gauss-Hermite Quadrature)**.** *The integrand part can be solved using Gauss-Hermite Quadrature. In numerical analysis, the method can be applied in the following form:*

$$\int_{-\infty}^\infty \exp(-x^2) f(x) dx \approx \sum_{i=1}^n \omega_i f(x_i) \tag{39}$$

where n is number of sample points used, and x_i is the roots of Hermite polnomial $H_n(x)$ such that $i = 1, ..., n$, and the weights ω_i is

$$\omega_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(x_i)]^2} \tag{40}$$

Applying Definition 1, the integral of the Laplace transform is then

$$\phi(s)^d = \frac{1}{\sqrt{\pi}} \sum_{q=1}^{N_q} \omega_q \exp(-s \exp(\sqrt{2}\sigma y_q)) \exp(\sqrt{2}\sigma y_q)^d \tag{41}$$

where q denotes the q-th element of Gauss Hermite Quadrature, i.e., ω_q denotes the q-th weight, y_q denotes the q-th node, and N_q denotes the total number of quadratures. Thus, substituting into the log-likelihood:

$$\ell_j(\boldsymbol{\theta}) = \sum_{i=1}^{n_j} \delta_{ij} \log(h(t_{ij}|\mathbf{x}_{ij})) + \log\Big(\frac{1}{\sqrt{\pi}} \sum_{q=1}^{N_q} \Bigg[\omega_q \exp(\sqrt{2}\sigma y_q)^{d_j} \exp\Big(- \sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) \exp(\sqrt{2}\sigma y_q)\Big)\Bigg]\Big) \tag{42}$$

Similarly, the ascertainment correction in the log-normal frailty can be written as

$$A_j(\boldsymbol{\theta}) = \Big[1 - \int_{-\infty}^{\infty} \exp(-zH(a_{p_j}|\mathbf{x}_{p_j}))f(z)dz\Big]^{I(T_{p_j}<a_{p_j})} \Big[\int_{-\infty}^{\infty} \exp(-zH(a_{p_j}|\mathbf{x}_{p_j}))f(z)dz\Big]^{1-I(T_{p_j}<a_{p_j})} \tag{43}$$

$$= \Big[1 - \sum_{q=1}^{N_q} \omega_q \exp\Bigg(-(\sum_{i=1}^{n_j} H(a_{p_j}|\mathbf{x}_{p_j})) \exp(\sqrt{2}\sigma y_{q_p})\Bigg)\Big]^{I(T_{p_j}<a_{p_j})} \tag{44}$$

$$\times \Big[\sum_{q=1}^{N_q} \omega_q \exp\Bigg(-(\sum_{i=1}^{n_j} H(a_{p_j}|\mathbf{x}_{p_j})) \exp(\sqrt{2}\sigma y_{q_p})\Bigg)\Big]^{1-I(T_{p_j}<a_{p_j})} \tag{45}$$

7. Likelihood and Missing Data

7.1. Reviews on Missing Data

In this subsection, the notations are **distinct** to all other sections or subsections. The missing data problem was firstly brought by Rubin [7], and further targetted as a major statistical problem which many methodologists have developed different statistical tools to handle the missing data. Such as the practical book written by Rubin [8], and some comprehensive reviews on current missing data problems by Baraldi and Enders [9]. The missing data mechanism was introduced by Little and Rubin [10]. There are three missing data mechanisms, which are Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). There are some reviews on the missing data which rigorously present the statistical concept of three types of the missing mechanism [11].

Definition 2. *(MCAR) Denote Y as the complete data matrix, and M as the missing data indicator matrix. Define y_{ij} and m_{ij} as i-th row (observation) and j-th column (variable) for the matrix Y and M. The conditional distribution of the missingness is said to be*

$$f(m_i|y_i, \phi) = f(m_i|\phi) \tag{46}$$

That is, for the parameters of this distribution, m_i does not depend on any observed or missing data.

Example 1. *(MCAR Example) There is a blind box with 500 indexed balls (No. 1 to 500) and their weights are unknown. We randomly draw 100 balls and measure their weights and record them in the Excel file. The Excel file contains two columns called Index and Weight, only those randomly selected balls will have Weights being filled. Those weights of unselected balls are called MCAR.*

Definition 3. *(MAR) Denote $y_{i,obs}$ as the observed y, and $y_{i,mis}$ as the missing y. Note that $y_i = (y_{i,obs}, y_{i,mis})$. The missing component is defined to be MAR if m only dependes on $y_{i,obs}$. That is,*

$$f(m_i|y_i, \phi) = f(m_i|y_{i,obs}, \phi) \tag{47}$$

Example 2. *(MAR Example) In a psychological study, participants are asked to complete a survey so the scientist can profile their personalities. One question that asks participants to report their Mood status being good or bad. Male participants are typically too shy to answer this question, which yields some responses being missing. This is called the MAR, that the missingness on Mood status depends on the participant's gender, but not on the missing Mood itself.*

Definition 4. *(MNAR) In the MNAR, the missingness depends on the missing data itself, which is*

$$f(m_i|y_i, \phi) = f(m_i|y_{i,mis}, y_{i,obs}, \phi) \tag{48}$$

In this case, the analysis needs to be conducted with caution. The missingness should be included in the likelihood construction.

Example 3. *(MNAR Example) There is a study on participants' incomes. Person A makes \$200,000 per year, so they decide to report this amount without hesitancies. Person B makes \$10,000 per year, so they are not willing to provide this information, which this response is left as blank. This type of missing depends on the missing data itself, that Person B refuses to provide the response due to the response being comparatively low.*

7.2. Without Considering the Kinship Structure

When assuming the data are missing at random, the missingness is only associated to the observed data. The frailty term and the missing data are therefore assumed independent. Denote $w_{ij} = (t_{ij}, \delta_{ij})$ be the observed survival data. From the complete log-likelihood:

$$\begin{aligned}
\ell_C(\boldsymbol{\theta}) &= \sum_{j=1}^{J}\sum_{i=1}^{n_j} \delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}, z_j) - H(t_{ij}|\mathbf{x}_{ij}, z_j) && (49)\\
&- \sum_{j=1}^{J} \log(1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}, z_j)) && (50)\\
&= \sum_{j=1}^{J}\sum_{i=1}^{n_j} \delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}) z_j - H(t_{ij}|\mathbf{x}_{ij}) z_j && (51)\\
&- \sum_{j=1}^{J} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) && (52)
\end{aligned}$$

In the MCEM framework, define $\boldsymbol{\theta}^{(r)}$ as r-th updates. The E-step can be written as

$$
\begin{aligned}
E(\ell_C(\boldsymbol{\theta})|\boldsymbol{\theta}^{(r)}) = & \sum_{j=1}^{J}\sum_{i=1}^{n_j}\int_{\mathbf{x}_{mis},z_j}\Big(\delta_{ij}\log h(t_{ij}|\mathbf{x}_{ij},z_j) - H(t_{ij}|\mathbf{x}_{ij},z_j)\Big) \tag{53}\\
& \times f(\mathbf{x}_{ij,mis},z_j|\mathbf{x}_{obs,ij},\boldsymbol{\theta}^{(r)})d\mathbf{x}_{ij,mis}dz_j \tag{54}\\
& -\sum_{j=1}^{J}\int_{\mathbf{x}_{mis},z_j}\log(1-\exp(z_jH_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \tag{55}\\
& \times f(\mathbf{x}_{ij,mis},z_j|\mathbf{x}_{obs,ij},\boldsymbol{\theta}^{(r)})d\mathbf{x}_{ij,mis}dz_j \tag{56}
\end{aligned}
$$

such that we need to integrate out the joint density of the frailty term and the missing covariate from $f(\mathbf{x}_{ij,mis},z_j|\mathbf{x}_{obs,ij},\boldsymbol{\theta}^{(r)})$. There are selections of the frailty density, such as Gamma distribution, log-normal distribution, and etc which have been discuseed in the previous chapter. In general, let's write $f(z_j|\upsilon)$ for the frailty distribution may be chosen with some parameters υ. Proposed by Herring et al. [3], the joint distribution of the frailty and the missing data can be adapted in our scenario that accounting for the ascertainment:

$$
\begin{aligned}
f(\mathbf{x}_{mis,ij},z_j|\mathbf{x}_{obs,ij},w_{ij},T_{p_j}<a_{p_j},\boldsymbol{\theta}^{(r)}) = & f(\mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij},w_{ij},z_j,T_{p_j}<a_{p_j},\boldsymbol{\theta}^{(r)}) \tag{57}\\
& \times f(z_j|\mathbf{x}_{obs,ij},w_{ij},T_{p_j}<a_{p_j},\boldsymbol{\theta}^{(r)}) \tag{58}
\end{aligned}
$$

we define $\Lambda = (\boldsymbol{\beta},\alpha,\lambda)$, then further we can write

$$
\begin{aligned}
& f(\mathbf{x}_{mis,ij},z_j|\mathbf{x}_{obs,ij},w_{ij},T_{p_j}<a_{p_j},\boldsymbol{\theta}^{(r)}) \tag{59}\\
& = \frac{f(w_{ij}|\mathbf{x}_{mis,ij},\mathbf{x}_{obs,ij},z_j,T_{p_j}<a_{p_j},\Lambda^{(r)})f(\mathbf{x}_{mis,ij},\mathbf{x}_{obs,ij}|\psi^{(r)})f(z_j|\upsilon^{(r)})}{\int_{\mathbf{x}_{mis,ij},z_j} f(w_{ij}|\mathbf{x}_{mis,ij},\mathbf{x}_{obs,ij},z_j,T_{p_j}<a_{p_j},\Lambda^{(r)})f(\mathbf{x}_{mis,ij},\mathbf{x}_{obs,ij}|\psi^{(r)})f(z_j|\upsilon^{(r)})dz_jd\mathbf{x}_{mis,ij}} \tag{60}\\
& \propto f(z_j|\upsilon^{(r)})\prod_{i=1}^{n_j} f(w_{ij}|\mathbf{x}_{mis,ij},\mathbf{x}_{obs,ij},z_j,T_{p_j}<a_{p_j},\Lambda^{(r)})f(\mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij},\psi^{(r)}) \tag{61}
\end{aligned}
$$

Clearly, we know $f(w_{ij}|\mathbf{x}_{mis,ij},\mathbf{x}_{obs,ij},z_j,T_{p_j}<a_{p_j},\boldsymbol{\beta}^{(r)})$ is the likelihood of one single observation i in family j, also we know the distribution of $f(\mathbf{x}_{ij}|\psi)$, as well as the frailty distribution $f(z_j|\upsilon)$. Therefore, in our case, we can apply Gibb's sampler where

1. We sample the missing data first, which we can obtain that

$$
\begin{aligned}
f(\mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij},w_{ij},z_j,T_{p_j}<a_{p_j},\boldsymbol{\theta}^{(r)}) \propto & f(w_{ij}|\mathbf{x}_{mis,ij},\mathbf{x}_{obs,ij},z_j,T_{p_j}<a_{p_j},\Lambda^{(r)}) \tag{62}\\
& \times f(\mathbf{x}_{mis,ij}|\mathbf{x}_{obs,ij},\psi^{(r)},T_{p_j}<a_{p_j}) \tag{63}\\
& \times P(T_{p_j}<a_{p_j}) \tag{64}
\end{aligned}
$$

In this case, the missing data will be filled for each iteration r, that being said, all data will be "observed" in this case. Therefore, we will use $\mathbf{x}_{ij}$ to simply denote completed covariates.

2. In order to approach the joint distribution, we now need to sample the frailty z_j from

$$f(z_j|\mathbf{x}_{ij}, w_{ij}, T_{p_j} < a_{p_j}, \boldsymbol{\theta}^{(r)}) \propto \prod_{i=1}^{n_j} f(w_{ij}|\mathbf{x}_{ij}, T_{p_j} < a_{p_j}, \Lambda^{(r)}) \times f(z_j|\upsilon^{(r)}) \tag{65}$$

3. The Gibb's sampler has been proven as an efficient sampling method to closely approach the desired joint distribution [12]. We can get a frailty distribution based on what we have sampled for the missing data, and we can obtain the missing data distribution based on what we have sampled from the frailty distribution.

These conditional densities can be explicitly written. We will obtain M completed dataset based on the Gibb's sampler. In general, without the specification of the frailty distribution, the E-step in MCEM can be written as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \Big(\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) - H(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) \Big) \tag{66}$$

$$+ \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \tag{67}$$

$$+ \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(\mathbf{x}_{ij,mis}^{(m)}, z_j^{(m)}|\mathbf{x}_{obs,ij}, \boldsymbol{\theta}) \tag{68}$$

$$= \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \Big(\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) - H(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) \Big) \tag{69}$$

$$+ \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \tag{70}$$

$$+ \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(\mathbf{x}_{mis,ij}^{(m)}|\mathbf{x}_{obs,ij}, \psi) + \sum_{j=1}^{J} \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(z_j^{(m)}|\upsilon) \tag{71}$$

Note that in this case, the distribution of the missing data is univariate since we are imputing each individual i in family j. However, in a genetic study, some missing covariates may need some considerations of the multivariate structure such as the kinship matrix.

7.3. Considering the Kinship Matrix

When we want to include the kinship matrix into the consideration, the distribution of the missing data becomes a multivariate distribution for family j. Denote $f(\mathbf{x}_{mis,j}|\mathbf{x}_{obs,j}, \boldsymbol{\psi})$ as the multivariate distribution of the missing data in family j. It is important to obtain the conditional distribution for each individual i conditioning on other individuals $-i$ within family j. Assume a n_j dimensional multivariate normal distribution of $\mathbf{x}_{mis} = (\mathbf{x}_{mis,1}, ..., \mathbf{x}_{mis,n_j})$ in family j, the index here will ignore the family index since if it's global, it will work too. The multivariate distribution will be assumed a normal, because we are focusing on the missing PRS which has a normal distributed behavior. The mean vector of this multivariate normal distribution can be written as $\boldsymbol{\mu} = (\mu_1, ..., \mu_{n_j})$ and the covariance matrix $\mathbf{\Sigma}$. Note that $\mathbf{\Sigma} = \tilde{\psi}_g^2 K + \tilde{\psi}_e^2$ and K is the kinship matrix with the diagonal of 1. $\tilde{\psi}_g^2$ is the genetic variance and $\tilde{\psi}_e^2$ is the residual variance. If we want to find the conditional distribution of each $x_{mis,i}$ in family j, given others X_{-i} where X_{-i} is the vector of all other vairables except $x_{mis,i}$. Partition the mean vector and the covariance matrix, suppose $\mathbf{X} = (x_i, X_{-i})$, then partition $\boldsymbol{\mu}$ and $\mathbf{\Sigma}$ accordingly:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_i \\ \boldsymbol{\mu}_{-i} \end{pmatrix} \tag{72}$$

and

$$\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{ii} & \mathbf{\Sigma}_{i,-i} \\ \mathbf{\Sigma}_{-i,i} & \mathbf{\Sigma}_{-i,-i} \end{pmatrix} \tag{73}$$

where Σ_{ii} is actually the variance of x_i. $\mathbf{\Sigma}_{i,-i}$ and $\mathbf{\Sigma}_{-i,i}$ are transpose of each other, and are covariances between x_i and X_{-i}. $\mathbf{\Sigma}_{-i,-i}$ is the covariance matrix of X_{-i}. Also, $\boldsymbol{\mu}$ can be estimated using a linear regression from a multivariate version with flexible covariance matrix introduced by Ziyatdinov et al. [13]. Then we can compute the conditional mean and variance from

$$E(x_i|X_{-i}) = \mu_i + \mathbf{\Sigma}_{i,-i}\mathbf{\Sigma}_{-i,-i}^{-1}(\mathbf{x}_{-i} - \boldsymbol{\mu}_{-i}) \tag{74}$$

and the variance

$$\mathrm{Var}(x_i|\mathbf{x}_{-i}) = \Sigma_{ii} - \mathbf{\Sigma}_{i,-i}\mathbf{\Sigma}_{-i,-i}^{-1}\mathbf{\Sigma}_{-i,i} \tag{75}$$

Once we can calculate these statistics, we are able to sample each x_i using a univariate normal distribution while still considering the kinship matrix. The MCEM will then perform using the same idea of the previous subsection.

8. Detailed Implementations of Gibb's Sampler

In section 7.2, the posterior distribution is not easy to sample from. So the Gibb's sampler needs an additional step called Matropolis-Hastings step. In the article by Herring et al. [3], the choice of the frailty distribution is multivariate normal, which satisfies the property of log-concavity for adaptive rejection algorithm. However, when the frailty distribution is designed to be Gamma or log-normal distribution, the log-concavity fails. One may use the Metropolis-Hastings (MH) step within Gibbs to determine the acceptance or rejection when sampling the posterior distributions [14]. There are some articles on the MH algorithm and how it works such as by Andrieu and Moulines [15], and some articles have discussed how Gibb's sampler are adapted using MCMC methods [16]. This section will discuss the MH-within-Gibbs algorithm when sampling the missing data and frailty. To sample the frailty using the Gibb's sampler with MH algorithm, the procedure is

1. We have a proposal sampling distribution $q(z_j'|z_j)$, which represents what we are sampling in the iteration r
2. We first sample from this $q(z_j'|z_j)$
3. For z_j, $q(z_j'|z_j) = f(z_j|\upsilon^{(r)})$

4. Then calculate the acceptance ratio:

$$\gamma = \min\left(1, \frac{\prod_{i=1}^{n_j} f(w_{ij}|\mathbf{x}_{ij}, z'_j, T_{p_j} < a_{p_j}, \Lambda^{(r)}) \times q(z'_j|z_j)}{\prod_{i=1}^{n_j} f(w_{ij}|\mathbf{x}_{ij}, z_j, T_{p_j} < a_{p_j}, \Lambda^{(r)}) \times q(z_j|z'_j)}\right) \tag{76}$$

5. Now we are sampling $\tilde{u} \sim \text{Unif}(0,1)$, and we accept z' if $\tilde{u} \leq \gamma$. Otherwise, reject and set $z'_j = z_j$.

The same idea for missing data. This MH-within-Gibbs is more flexible when log-concavity does not hold for the posterior distribution, also when it's impossible to direct sample from the posterior distribtion.

9. Multiple Imputation

In the MCEM, the likelihood is explicitly written out and it "automatically" solves the missing data and complete data within an iterative process to make the inference of model parameters. However, there are several disadvantages such as a highly costly computational time, since parameters are estimated all at once based on the likelihood equation. Moreover, there are arguments that during the MCEM, the precision is questionable because there is no distinction between the observed and imputed data, although their weights are evaluated in the E-step since the integral was broken down to a summation [17]. There are also selections of multiple imputation methods, which the most common way is to use the linear regression using Bayesian framework to make imputations on continuous variables, same idea of using logistic regression for binary variables [8].

9.1. Multiple Imputation for the Continuous Variable without Considering the Family Structure

Again, to avoid overloaded mathematical notations, this subsection will not follow the previously defined notations. When making the imputation on the continuous variable, one easy way is to assume a conditionally normal model. Suppose y is the variable contains missing values, $\mathbf{x}$ are other variables are fully observed. We assume there are total p parameters being estimated in this linear regression. The conditionally normal model (linear regression) can be written as

$$y|\mathbf{x}, \boldsymbol{\beta} \sim N(\boldsymbol{\beta}\mathbf{x}, \sigma^2) \tag{77}$$

In the linear regression setting,

$$y = \mathbf{x}\boldsymbol{\beta} + \epsilon,\ \ \epsilon \sim N(0, \sigma^2) \tag{78}$$

This simply corresponds to the likelihood

$$f(y|\mathbf{x}, \boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-n/2} \exp\Big(-\frac{1}{2\sigma^2}(y - \mathbf{x}\boldsymbol{\beta})^\top (y - \mathbf{x}\boldsymbol{\beta})\Big) \tag{79}$$

because of the conditional normality of y. We can solve that

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top y \tag{80}$$

In the Bayesian framework, we need to find the prior, which is the joint distribution $f(\sigma^2, \boldsymbol{\beta})$. Note that $(y - \mathbf{x}\boldsymbol{\beta})^\top(y - \mathbf{x}\boldsymbol{\beta})$ can be written as

$$\begin{aligned}(y - \mathbf{x}\boldsymbol{\beta})^\top(y - \mathbf{x}\boldsymbol{\beta}) &= \big((y - \mathbf{x}\boldsymbol{\beta}) + (\mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{x}\boldsymbol{\beta})\big)^\top\big((y - \mathbf{x}\boldsymbol{\beta}) + (\mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{x}\boldsymbol{\beta})\big) & (81)\\ &= (y - \mathbf{x}\boldsymbol{\beta})^\top(y - \mathbf{x}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top(\mathbf{x}^\top\mathbf{x})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + 2(\mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{x}\boldsymbol{\beta}) & (82)\\ &= (y - \mathbf{x}\boldsymbol{\beta})^\top(y - \mathbf{x}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top(\mathbf{x}^\top\mathbf{x})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) & (83)\end{aligned}$$

So Equation 79 will become

$$f(y|\mathbf{x}, \boldsymbol{\beta}, \sigma^2) \propto \underbrace{(\sigma^2)^{-v/2}\exp(-\frac{vs^2}{2\sigma^2})}_{f(\sigma)}\underbrace{(\sigma^2)^{-\frac{n-v}{2}}\exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top(\mathbf{x}^\top\mathbf{x})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\Big)}_{f(\boldsymbol{\beta}|\sigma)} \tag{84}$$

where $vs^2 = (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})^\top(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}}) = SSE$, such that $v = n_{obs} - p$. Then $f(\sigma^2)$ can be written as a proportional density to the inverse gamma distribution,

$$f(\sigma^2) \propto (\sigma^2)^{-\frac{v}{2}-1}\exp(-\frac{vs^2}{2\sigma^2}) = (\sigma^2)^{-\frac{v}{2}-1}\exp(-\frac{SSE}{2\sigma^2}) \tag{85}$$

In this $\sigma^2 \sim$ Inverse-Gamma(α, ϕ), we have $\alpha = \frac{v}{2} = \frac{n_{obs}-p}{2}$ and $\phi = \frac{1}{2}vs^2 = \frac{SSE}{2}$. From the inverse-gamma property, when

$$\sigma^2 \sim \text{Inverse-Gamma}(\frac{n_{obs} - p}{2}, \frac{SSE}{2}) \tag{86}$$

and $\exists\lambda$ such that

$$\sigma^2 = \frac{SSE}{2}/\lambda \tag{87}$$

then λ can be transformed

$$\lambda = \frac{\chi^2_{n_{obs}-p}}{2} \tag{88}$$

since $\chi^2_{df} = \text{Gamma}(\frac{df}{2}, 2)$, so

$$\sigma^2 = \frac{\frac{SSE}{2}}{\frac{\chi^2_{n_{obs}-p}}{2}} = \frac{SSE}{\chi^2_{n_{obs}-p}} \tag{89}$$

Thus, we can sample the standard deviation of the missing data distribution from

$$\sigma^* = \hat{\sigma}\sqrt{\frac{SSE}{\chi^2_{n_{obs}-p}}} = \hat{\sigma}\sqrt{\frac{SSE}{g}} \tag{90}$$

from sampling $g \sim \chi^2_{n_{obs}-p}$. Moreover, we know

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{x}^\top\mathbf{x}^{-1}) = \mathbf{V} \tag{91}$$

so the marginal distribution for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} \sim N(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{x}^\top\mathbf{x})^{-1}) \tag{92}$$

Note that when a random variable $T \sim N(\mathbf{m}, \mathbf{c})$, then T can be generated from a standard normal variable $\mathbf{u}$ with

$$T = \mathbf{m} + \mathbf{L}\mathbf{u} \tag{93}$$

where $\mathbf{L}$ is the cholesky decomposition of $\mathbf{c}$ such that $\mathbf{c} = \mathbf{L}\mathbf{L}^\top$. For $\boldsymbol{\beta}$, from 92, it can be derived as

$$\begin{align}\boldsymbol{\beta} &= \hat{\boldsymbol{\beta}} + \sigma(\mathbf{x}^\top\mathbf{x})^{-1/2}\mathbf{u}_1 \tag{94}\\ &= \hat{\boldsymbol{\beta}} + \mathbf{u}_1\mathbf{V}^{-1/2} \tag{95}\end{align}$$

Adjusting for σ^* to make sure $\boldsymbol{\beta}$ matches the variability implied by the random draw of σ^*,

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u}_1\mathbf{V}^{-1/2} \tag{96}$$

such that $\mathbf{u}_1$ is a row vector of p independent draws from a standard normal distribution, $u_{1k} \overset{iid}{\sim} N(0,1)$, and $k = 1, ..., p$. The imputation for y_i^* is computed as

$$y_i^* = \boldsymbol{\beta}^*\mathbf{x}_i + u_{2i}\sigma^*, \ s.t. \ u_{2i} \sim N(0,1) \tag{97}$$

where u_{2i} adds the uncertainty to the imputation as well to ensure the imputation is not solely based on the predicted value of y_i^*. This prevents the underestimation of the variability. Therefore, the comprehensive steps of the multiple imputation on the continuous missing data can be summarized to the following steps:

1. Calculate $\hat{y} = \hat{\boldsymbol{\beta}}\mathbf{x}$ using y_{obs}, and $\hat{\boldsymbol{\beta}}$ can be obtained easily, as well as $\hat{\sigma}$, and $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{V}$
2. Draw $g \sim \chi^2_{n_{obs}-p}$ for one random draw
3. Calculate $\sigma^* = \hat{\sigma}/\sqrt{SSE/g}$
4. Draw a p dimensional vector $\mathbf{u}_1$ such that $u_{1k} \overset{iid}{\sim} N(0,1)$ and $k = 1, ..., p$
5. Calculate $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \frac{\sigma^*}{\hat{\sigma}}\mathbf{u}_1\mathbf{V}^{1/2}$ such that $\mathbf{V}^{1/2}$ is the cholesky decomposition of $\mathbf{V}$
6. Draw $u_{2i} \overset{iid}{\sim} N(0,1)$
7. Impute $y_{mis,i} = \boldsymbol{\beta}^*\mathbf{x}_i + u_{2i}\sigma^*$
8. Repeat 2. to 7. for M times to obtain M complete datasets

9.2. Multiple Imputation for the Continuous Variable Considering the Family Structure

In order to account for the kinship correlations, the conditional normal distribution needs to be adjusted where the variable contains the missing components are said to be multivariate

$$\mathbf{y}|\mathbf{x}, \boldsymbol{\beta} \sim MVN(\boldsymbol{\beta}\mathbf{x}, \sigma_g^2 K + \sigma_e^2) \tag{98}$$

where in the linear mixed effect regression form with flexible covariance matrix, the model can be written as

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e} \tag{99}$$

where $\mathbf{u} \sim MVN(0, \sigma_g^2 K)$, and $\mathbf{e} \sim MVN(0, \sigma_e^2 I)$, such that K is the kinship matrix with the diagonal of 1 and I is the identity matrix. Denote that $\boldsymbol{\Sigma} = \sigma_g^2 K + \sigma_e^2$. So in this multivariate version of linear mixed effects model. With unknown mean and covariance matrix, the prior of the covariance matrix can be selected as an Inverse Wishart distribution with some degrees of freedom and the scale parameter. However, in this case, the covariance matrix is not fully unknown. Therefore, some steps of this multiple imputation will be based on the empirical estimates rather than the prior distributions. The step can be concluded as follows:

1. Obtain the kinship matrix K among all individuals
2. Calculate the estimates of $\hat{y} = \mathbf{x}\boldsymbol{\beta}$, obtain estimates of $\hat{\boldsymbol{\beta}}$, $\hat{\sigma_g}^2$, $\hat{\sigma_e}^2$, $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{V}$. In this step, naturally, $\hat{\boldsymbol{\Sigma}}$ is obtained.
3. Obtain the conditional variance of y_i,

$$\mathrm{Var}(y_i|\mathbf{y}_{-i}) = \hat{\Sigma}_{ii} - \hat{\boldsymbol{\Sigma}}_{i,-i}\hat{\boldsymbol{\Sigma}}_{-i,-i}^{-1}\hat{\boldsymbol{\Sigma}}_{-i,i} = \hat{\sigma_i}^2 \tag{100}$$

4. Draw p-dimensional vector w_1 such that $w_{1k} \overset{iid}{\sim} N(0,1)$ where $k = 1, ..., p$
5. Calculate $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + w_1 \mathbf{V}^{1/2}$ such that $\mathbf{V}^{1/2}$ is the cholesky decomposition of $\mathbf{V}$
6. Obtain $\mu_i^* = \boldsymbol{\beta}^* \mathbf{x}_i$
7. Obtain the conditional expectations

$$E(y_{mis,i}|\mathbf{y}_{-i}) = \mu_i^* + \hat{\boldsymbol{\Sigma}}_{i,-i}\hat{\boldsymbol{\Sigma}}_{-i,-i}^{-1}(\mathbf{y}_{-i} - \boldsymbol{\mu}_{-i}^*) \tag{101}$$

8. Draw $w_{2i} \sim N(0,1)$
9. Impute $y_{mis,i} = E(y_{mis,i}|\mathbf{y}_{-i}) + w_{2i}\hat{\sigma}_i$
10. Repeat 4. to 9. for M times to obtain M complete datasets.

10. Simulation Study (Temporary)

10.1. Generating Missing at Random

From the MICE (Multivariate Imputation by Chained Equations) package in R authored by Van Buuren and Groothuis-Oudshoorn [18], there is one function add-on that is very helpful in generating the missingness of a data while considering the missing mechanism. The function was introduced by Schouten et al. [19] that on the opposite of imputation, the amputation is designed to simulate the missing data.

References

[1] Jacqueline E Rudolph, Stephen R Cole, and Jessie K Edwards. Parametric assumptions equate to hidden observations: comparing the efficiency of nonparametric and parametric models for estimating time to aids or death in a cohort of hiv-positive women. *BMC medical research methodology*, 18:1–5, 2018.

[2] Roger L Berger and George Casella. *Statistical inference*. Duxbury, 2001.

[3] Amy H Herring, Joseph G Ibrahim, and Stuart R Lipsitz. Frailty models with missing covariates. *Biometrics*, 58(1):98–109, 2002.

[4] Suyeon Park, Sungyoung Lee, Young Lee, Christine Herold, Basavaraj Hooli, Kristina Mullin, Taesung Park, Changsoon Park, Lars Bertram, Christoph Lange, et al. Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families. *BMC medical genetics*, 16:1–12, 2015.

[5] Andrew G Clark, Melissa J Hubisz, Carlos D Bustamante, Scott H Williamson, and Rasmus Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome research*, 15(11):1496–1502, 2005.

[6] Theodor A Balan and Hein Putter. A tutorial on frailty models. *Statistical methods in medical research*, 29(11):3424–3454, 2020.

[7] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[8] Donald B Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, New York, 1987. doi: 10.1002/9780470316696.

[9] Amanda N Baraldi and Craig K Enders. An introduction to modern missing data analyses. *Journal of school psychology*, 48(1):5–37, 2010.

[10] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[11] Miriam Seoane Santos, Ricardo Cardoso Pereira, Adriana Fonseca Costa, Jastin Pompeu Soares, João Santos, and Pedro Henriques Abreu. Generating synthetic missing data: A review by missing mechanism. *IEEE Access*, 7:11651–11667, 2019.

[12] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[13] Andrey Ziyatdinov, Miquel Vázquez-Santiago, Helena Brunel, Angel Martinez-Perez, Hugues Aschard, and Jose Manuel Soria. lme4qtl: linear mixed models with flexible covariance structure for genetic studies of related individuals. *BMC bioinformatics*, 19: 1–5, 2018.

[14] Jim E Griffin and Stephen G Walker. On adaptive metropolis–hastings methods. *Statistics and Computing*, 23:123–134, 2013.

[15] Christophe Andrieu and Éric Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3):1462 – 1505, 2006.

[16] Krzysztof Łatuszyński, Gareth O. Roberts, and Jeffrey S. Rosenthal. Adaptive Gibbs samplers and related MCMC methods. *The Annals of Applied Probability*, 23(1):66 – 98, 2013.

[17] Stef Van Buuren. Multiple imputation of multilevel data. In *Handbook of advanced multilevel analysis*, pages 173–196. Routledge, 2011.

[18] Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

[19] Rianne Margaretha Schouten, Peter Lugtig, and Gerko Vink. Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, 88(15):2909–2930, 2018.