

Jiaqi's Thesis Progress Report (Updated Mar. 29)

Jiaqi Bi^a

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

1. To Do List

1. MCEM - Both univariate and multivariate missing data
2. Add a section that talk about when dealing with missing data without using joint distribution of frailty and missing covariate

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
n	Number of individuals
J	Number of Families (Clusters)
z	Frailty term
$h(\cdot)$	Hazard fuction
$h_0(\cdot)$	Baseline hazard function
$H(\cdot)$	Cumulative hazard fuction
$S(\cdot)$	Survival fuction
$A_j(\cdot)$	Ascertainment of family j into the study
\mathbf{x}	Covariates
β	Model coefficients vector
θ	Parameter vector
λ	Weibull shape parameter
α	Weibull scale parameter
k	Gamma shape and rate parameters

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

σ^2 Log-Normal variance parameter

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

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

There are total n_j individuals in family j , where $i = 1, \dots, n_j$, and total J families that $j = 1, \dots, J$. $x_{1,ij}$ 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

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

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

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

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

$$H(t_{ij}|\mathbf{x}_{ij}, z_j) = \int_0^t h_{ij}(u|\mathbf{x}_{ij}, g_{ij}, z_j) du \quad (4)$$

$$= \alpha \xi_{ij} z_j \lambda \int_0^t u^{\lambda-1} du \quad (5)$$

$$= \alpha \xi_{ij} z_j \lambda \cdot \frac{1}{\lambda} t_{ij}^\lambda = \alpha \xi_{ij} z_j t_{ij}^\lambda \quad (6)$$

and the survival function

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

Let $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \alpha, \lambda, v\}$, where v 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

$$L(\boldsymbol{\theta}) = \prod_{j=1}^J \prod_{i=1}^{n_j} (\alpha \lambda t_{ij}^{\lambda-1} \xi_{ij} z_j)^{\delta_{ij}} \exp(-\alpha \xi_{ij} z_j t_{ij}^\lambda) \quad (8)$$

$$= \prod_{j=1}^J \prod_{i=1}^{n_j} h(t_{ij}|\mathbf{x}_{ij}, z_j)^{\delta_{ij}} \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) \quad (9)$$

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

$$L(\boldsymbol{\theta}) = \prod_{j=1}^J \prod_{i=1}^{n_j} \int_{z_j} (\alpha \lambda t_{ij}^{\lambda-1} \xi_{ij} z_j)^{\delta_{ij}} \exp(-\alpha \xi_{ij} z_j t_{ij}^{\lambda}) f(z_j) dz_j \quad (10)$$

$$= \prod_{j=1}^J \prod_{i=1}^{n_j} \int_{z_j} h(t_{ij} | \mathbf{x}_{ij}, z_j)^{\delta_{ij}} \exp(-H(t_{ij} | \mathbf{x}_{ij}, z_j)) f(z_j) dz_j \quad (11)$$

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].

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

$$= \prod_{j=1}^J \prod_{i=1}^{n_j} \int_{z_j, \mathbf{x}_{mis,ij}} h(t_{ij} | \mathbf{x}_{ij}, z_j)^{\delta_{ij}} \exp(-H(t_{ij} | \mathbf{x}_{ij}, z_j)) f(z_j, \mathbf{x}_{mis,ij}) dz_j d\mathbf{x}_{mis,ij} \quad (13)$$

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

23 4. Ascertainment Correction

24 Within a genetic study, those families are typically selected when there is an affected
 25 person called a proband. This will yield a selection bias because this is no long a case-control
 26 study, and can potentially defect the statistical power [4, 5]. It is crucial to address the
 27 ascertainment bias. For each family j , the ascertainment A_j is defined to be the probability
 28 of the proband p being ascertained by the age a_{p_j} at examination, i.e., $A_j = P(T_{p_j} < a_{p_j})$
 29 where a_{p_j} is proband's age at study entry. Applying the ascertainment correction for the
 30 log-likelihood in family j :

$$\tilde{\ell}_j(\boldsymbol{\theta}) = \ell_j(\boldsymbol{\theta}) - \log A_j(\boldsymbol{\theta}) \quad (14)$$

31 where $\tilde{\ell}$ is the log-likelihood with ascertainment correction, and ℓ is the crude log-likelihood.
 32 Define \mathbf{x}_{p_j} the covariate matrix for proband in family j .

33 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 $\mathcal{L}(f(z)) = \phi(s)$ where $s = \sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij})$:

$$\phi(s) = \int_0^\infty e^{-sz} f(z) dz \quad (15)$$

$$= \int_0^\infty e^{-sz} \frac{k^k}{\Gamma(k)} z^{k-1} e^{-kz} dz \quad (16)$$

34 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

$$\phi(s) = \frac{k^k}{\Gamma(k)} \int_0^\infty e^{-(s+k)z} z^{k-1} dz = \frac{k^k}{\Gamma(k)} \cdot \frac{\Gamma(k)}{(s+k)^k} = \left(1 + \frac{s}{k}\right)^{-k} \quad (17)$$

35 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 $\phi(s)^{(d)}$:

$$\phi(s)^{(d)} = (-1)^d \int_0^\infty z^d e^{-sz} f(z) dz \quad (18)$$

$$= (-1)^d \frac{(k+d-1)!}{(k-1)!(s+k)^d} \left(1 + \frac{s}{k}\right)^{-k} \quad (19)$$

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

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^k \log \left[\int_0^\infty \prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}, z_j))^{\delta_{ij}} \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) f(z_j) dz_j \right] \quad (20)$$

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

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

$$= \sum_{j=1}^J \log \left[\prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \frac{(k+d_j-1)!}{(k-1)!(\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) + k)^{d_j}} \left(1 + \frac{\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij})}{k}\right)^{-k} \right] \quad (23)$$

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

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

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

Note we can still apply Laplace transform for the ascertainment correction, such that

$$A_j(\boldsymbol{\theta}) = 1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}) \quad (27)$$

$$= 1 - \int_0^\infty S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}, z_j) f(z_j) dz_j \quad (28)$$

$$= 1 - \int_0^\infty \exp(-z_j \cdot H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})) f(z_j) dz_j \quad (29)$$

$$= 1 - \left(1 + \frac{H_{p_j}(a_{p_j}|\mathbf{x}_{p_j})}{k}\right)^{-k} \quad (30)$$

36 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

$$E(\log Z) = 0 \quad (31)$$

$$\text{Var}(\log Z) = \sigma^2 \quad (32)$$

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

$$f(z) = \frac{1}{\sqrt{2\pi}\sigma} z^{-1} \exp\left(-\frac{\log(z)^2}{2\sigma^2}\right) \quad (33)$$

38 The Laplace transform is then

$$\phi(s) = \mathcal{L}(f_Z)(s) = \int_0^\infty \exp(-sz) \cdot f(z) dz \quad (34)$$

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:

$$\phi(s)^d = \int_{-\infty}^\infty z^d \exp(-sz) \cdot \frac{1}{\exp(\sqrt{2}\sigma y) \sigma \sqrt{2\pi}} \cdot \exp(-y^2) \cdot \sqrt{2}\sigma \exp(\sqrt{2}\sigma y) dy \quad (35)$$

$$= \int_{-\infty}^\infty \exp(\sqrt{2}\sigma y)^d \exp(-s \exp(\sqrt{2}\sigma y)) \cdot \frac{1}{\sqrt{\pi}} \exp(-y^2) dy \quad (36)$$

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

$$\int_{-\infty}^\infty \exp(-x^2) f(x) dx \approx \sum_{i=1}^n \omega_i f(x_i) \quad (37)$$

41 where n is number of sample points used, and x_i is the roots of Hermite polynomial $H_n(x)$

such that $i = 1, \dots, n$, and the weights ω_i is

$$\omega_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(x_i)]^2} \quad (38)$$

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

$$\phi(s)^d = \frac{1}{\sqrt{\pi}} \sum_{q=1}^{N_q} \omega_q \exp(-s \exp(\sqrt{2}\sigma y_q)) \exp(\sqrt{2}\sigma y_q)^d \quad (39)$$

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:

$$\ell_j(\boldsymbol{\theta}) = \sum_{i=1}^{n_j} \delta_{ij} \log(h(t_{ij}|\mathbf{x}_{ij})) + \log \left(\frac{1}{\sqrt{\pi}} \sum_{q=1}^{N_q} \left[\omega_q \exp(\sqrt{2}\sigma y_q)^{d_j} \exp \left(- \sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) \exp(\sqrt{2}\sigma y_q) \right) \right] \right) \quad (40)$$

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

$$A_j(\boldsymbol{\theta}) = 1 - \int_{-\infty}^{\infty} \exp(-z H(a_{p_j}|\mathbf{x}_{p_j})) f(z) dz \quad (41)$$

$$= 1 - \sum_{q=1}^{N_q} \omega_q \exp \left(- \left(\sum_{i=1}^{n_j} H(a_{p_j}|\mathbf{x}_{p_j}) \exp(\sqrt{2}\sigma y_{q_p}) \right) \right) \quad (42)$$

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

$$f(m_i|y_i, \phi) = f(m_i|\phi) \quad (43)$$

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 depends on $y_{i,obs}$. That is,

$$f(m_i|y_i, \phi) = f(m_i|y_{i,obs}, \phi) \quad (44)$$

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

$$f(m_i|y_i, \phi) = f(m_i|y_{i,mis}, y_{i,obs}, \phi) \quad (45)$$

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 hesitations. 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. Log-Likelihood and Missing at Random

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:

$$\ell_C(\boldsymbol{\theta}) = \sum_{j=1}^J \sum_{i=1}^{n_j} \delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}, z_j) - H(t_{ij}|\mathbf{x}_{ij}, z_j) \quad (46)$$

$$- \sum_{j=1}^J \log(1 - S_{p_j}(a_{p_j}|\mathbf{x}_{p_j}, z_j)) \quad (47)$$

$$= \sum_{j=1}^J \sum_{i=1}^{n_j} \delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}) z_j - H(t_{ij}|\mathbf{x}_{ij}) z_j \quad (48)$$

$$- \sum_{j=1}^{n_j} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \quad (49)$$

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

$$E(\ell_C(\boldsymbol{\theta})|\boldsymbol{\theta}^{(r)}) = \sum_{j=1}^J \sum_{i=1}^{n_j} \int_{\mathbf{x}_{mis}, z_j} \left(\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}, z_j) - H(t_{ij}|\mathbf{x}_{ij}, z_j) \right) \quad (50)$$

$$\times f(\mathbf{x}_{ij,mis}, z_j | \mathbf{x}_{obs,ij}, \boldsymbol{\theta}^{(r)}) d\mathbf{x}_{ij,mis} dz_j \quad (51)$$

$$- \sum_{j=1}^J \int_{\mathbf{x}_{mis}, z_j} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \quad (52)$$

$$\times f(\mathbf{x}_{ij,mis}, z_j | \mathbf{x}_{obs,ij}, \boldsymbol{\theta}^{(r)}) d\mathbf{x}_{ij,mis} dz_j \quad (53)$$

91 such that we need to integrate out the joint density of the frailty term and the missing
 92 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
 93 Gamma distribution, log-normal distribution, and etc. In general, let's write $f(z_j|\nu)$ for the
 94 frailty distribution may be chosen with some parameters ν . Proposed by Herring et al. [3],
 95 the joint distribution of the frailty and the missing data can be adapted in our scenario that
 96 accounting for the ascertainment:

$$f(\mathbf{x}_{mis,ij}, z_j | \mathbf{x}_{obs,ij}, w_{ij}, T_{p_j} < a_{p_j}, \boldsymbol{\theta}^{(r)}) = f(\mathbf{x}_{mis,ij} | \mathbf{x}_{obs,ij}, w_{ij}, z_j, T_{p_j} < a_{p_j}, \boldsymbol{\theta}^{(r)}) \quad (54)$$

$$\times f(z_j | \mathbf{x}_{obs,ij}, w_{ij}, T_{p_j} < a_{p_j}, \boldsymbol{\theta}^{(r)}) \quad (55)$$

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

$$f(\mathbf{x}_{mis,ij}, z_j | \mathbf{x}_{obs,ij}, w_{ij}, T_{p_j} < a_{p_j}, \boldsymbol{\theta}^{(r)}) \quad (56)$$

$$= \frac{f(w_{ij} | \mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij}, z_j, T_{p_j} < a_{p_j}, \Lambda^{(r)}) f(\mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij} | \psi^{(r)}) f(z_j | v^{(r)})}{\int_{\mathbf{x}_{mis,ij}, z_j} f(w_{ij} | \mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij}, z_j, T_{p_j} < a_{p_j}, \Lambda^{(r)}) f(\mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij} | \psi^{(r)}) f(z_j | v^{(r)}) dz_j d\mathbf{x}_{mis,ij}} \quad (57)$$

$$\propto f(w_{ij} | \mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij}, z_j, T_{p_j} < a_{p_j}, \Lambda^{(r)}) f(\mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij} | \psi^{(r)}) f(z_j | v^{(r)}) \quad (58)$$

98 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
 99 observation i in family j , also we know the distribution of $f(\mathbf{x}_{ij} | \psi)$, as well as the frailty
 100 distribution $f(z_j | v)$. Therefore, in our case, we can apply Gibb's sampler where

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

$$f(\mathbf{x}_{mis,ij} | \mathbf{x}_{obs,ij}, w_{ij}, z_j, T_{p_j} < a_{p_j}, \boldsymbol{\theta}^{(r)}) \propto f(w_{ij} | \mathbf{x}_{mis,ij}, \mathbf{x}_{obs,ij}, z_j, T_{p_j} < a_{p_j}, \Lambda^{(r)}) \quad (59)$$

$$\times f(\mathbf{x}_{mis,ij} | \mathbf{x}_{obs,ij}, \psi^{(r)}) \quad (60)$$

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

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

$$f(z_j | \tilde{\mathbf{x}}_{ij}, w_{ij}, T_{p_j} < a_{p_j}, \boldsymbol{\theta}^{(r)}) \propto \prod_{i=1}^{n_j} f(w_{ij} | \tilde{\mathbf{x}}_{ij}, T_{p_j} < a_{p_j}, \Lambda^{(r)}) \times f(z_j | v^{(r)}) \quad (61)$$

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

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

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \sum_{j=1}^J \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \left(\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) - H(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) \right) \quad (62)$$

$$+ \sum_{j=1}^J \frac{1}{M_j} \sum_{m=1}^{M_j} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \quad (63)$$

$$+ \sum_{j=1}^J \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(\mathbf{x}_{ij,mis}^{(m)}, z_j^{(m)}|\mathbf{x}_{obs,ij}, \boldsymbol{\theta}) \quad (64)$$

$$= \sum_{j=1}^J \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \left(\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) - H(t_{ij}|\mathbf{x}_{ij}^{(m)}, z_j^{(m)}) \right) \quad (65)$$

$$+ \sum_{j=1}^J \frac{1}{M_j} \sum_{m=1}^{M_j} \log(1 - \exp(z_j H_{p_j}(a_{p_j}|\mathbf{x}_{p_j}))) \quad (66)$$

$$+ \sum_{j=1}^J \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(\mathbf{x}_{mis,ij}^{(m)}|\mathbf{x}_{obs,ij}, \psi) + \sum_{j=1}^J \frac{1}{M_j} \sum_{m=1}^{M_j} \sum_{i=1}^{n_j} \log f(z_j^{(m)}|v) \quad (67)$$

where for example, when $x_{mis,ij,1}$ is missing and $x_{obs,ij,2}$ is observed, we take

$$f(x_{mis,ij,1}^{(m)}|x_{obs,ij,2}, \psi) = \frac{1}{\tilde{\psi}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x_{mis,ij,1}^{(m)} - \mu}{\tilde{\psi}}\right)^2\right) \quad (68)$$

such that $\hat{\mu} = \psi_0 + \psi_1 x_{obs,ij,2}$, and $\tilde{\psi}$ is the standard error of the conditional normal distribution. We can apply the kinship matrix for a family study, that we will deal with a multivariate normal distributed $\mathbf{x}_{mis,j,1}^{(m)}$ for family j .

$$f(\mathbf{x}_{mis,j,1}^{(m)}|\mathbf{x}_{obs,j,2}, \psi, K) = (2\pi)^{(-\frac{n_j}{2})} \det(\Sigma)^{(-\frac{1}{2})} \exp\left(-\frac{1}{2}(\mathbf{x}_{mis,j,1}^{(m)} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_{mis,j,1}^{(m)} - \boldsymbol{\mu})\right) \quad (69)$$

where $\Sigma = \tilde{\psi}_g^2 K + \tilde{\psi}_e^2$, and K is the kinship matrix with diagonals of 1 and is known. Also, $\hat{\boldsymbol{\mu}} = \boldsymbol{\psi}_0 + \boldsymbol{\psi}_1 \mathbf{x}_{obs,j,2}$. Therefore, we may sample from $\mathbf{x}_{mis,j,1} \sim MVN(\boldsymbol{\mu}, \tilde{\psi}_g^2 K + \tilde{\psi}_e^2)$. Moreover, we can give the specific distribution of the frailty in the E-step by sampling the frailty term z_j , for example, when z_j is assumed to be Gamma distribution, then

$$f(z_j^{(m)}|k) = \frac{k^k}{\Gamma(k)} (z_j^{(m)})^{k-1} \exp(-k z_j^{(m)}) \quad (70)$$

115 When z_j is log-normal distributed with the mean 0, we have

$$f(z_j^{(m)}|v) = \frac{1}{z_j^{(m)}v\sqrt{2\pi}} \exp\left(-\frac{(\log z_j^{(m)})^2}{2v^2}\right) \quad (71)$$

116 Thus, the $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$ function will yield a closed form that can be optimized via M-step, with
 117 the iterations that update all parameters using Gibb's sampler.

118 8. Missing PRS and MCEM

119 8.1. Not Considering the Family Correlations

We propose an MCEM framework in terms of estimating the distribution of the PRS. The PRS was calculated to infer the relationship between a phenotype and multiple genetic loci, while these information were not gained if one was not involved in the original GWAS. Thus, we propose to sample the PRS using the information that we have already obtained through the study. Denote $\mathbf{x}_{j,1}$ as the PRS scores vector in family j , and $\mathbf{x}_{j,2}$ the mutation status vector in family j . Take \mathbf{p}_j as the proband indicator vector in family j , \mathbf{c}_j is the current age for patients in family j . So we now have a design matrix when modelling the missing PRS, call it $\mathbf{W} = (\log(\mathbf{t}_j), \boldsymbol{\delta}_j, \log(\mathbf{t}_j) \odot \boldsymbol{\delta}_j, \mathbf{p}_j, \mathbf{c}_j, \mathbf{x}_{j,2})$. We can make the assumption on the conditional distribution of the PRS, take $\mathbf{X}_{j,1}|\mathbf{W} \sim MVN(\mathbf{W}\boldsymbol{\psi} + \mathbf{u}, \sigma^2\mathbf{I})$. We are interested in modelling the PRS while accounting for the between family variance, so $\mathbf{u} \sim MVN(0, \sigma_u^2\mathbf{I})$. Thus, the E-step for Gamma frailty model with ascertainment correction is then

$$E_{\mathbf{x}_{j,1,mis}}(\ell(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})) = \sum_{j=1}^J \left[\sum_{i=1}^{n_j} \int_{\mathbf{x}_{j,1,mis}} (\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij})) + \right. \quad (72)$$

$$+ \log\left(\frac{(k+d_j-1)!}{k!k^{d_j-1}} \left(1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}))}{k}\right)^{-k-d_j}\right) - \quad (73)$$

$$\left. - \log(A_j(\boldsymbol{\theta})) + \log f(x_{ij,1,mis}|w_{ij}, \boldsymbol{\psi}) dx_{ij,1,mis} \right] \quad (74)$$

Taking a sample of size M when we sample $f(x_{ij,1,mis}|w_{ij}, \boldsymbol{\psi})$ for each subject i in family j , $(x_{ij,i,mis}^{(1)}, \dots, x_{ij,i,mis}^{(M)})$. This leads to an E-step:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \sum_{j=1}^J \left[\sum_{i=1}^{n_j} \sum_{m=1}^M (\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}^{(m)})) + \right. \quad (75)$$

$$+ \log\left(\frac{(k+d_j-1)!}{k!k^{d_j-1}} \left(1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}^{(m)}))}{k}\right)^{-k-d_j}\right) - \quad (76)$$

$$\left. - \log(A_j(\boldsymbol{\theta})) + \log f(x_{ij,1,mis}^{(m)}|w_{ij}, \boldsymbol{\psi}) \right] \quad (77)$$

Similarly, the expectation with respect to the missing PRS in log-normal frailty can be written as

$$E_{\mathbf{X}_{j,1,mis}}(\ell(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})) = \sum_{j=1}^J \left[\sum_{i=1}^{n_j} \int_{\mathbf{X}_{j,1,mis}} (\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij})) + \right. \quad (78)$$

$$+ \log \left(\frac{1}{\sqrt{\pi}} \sum_{q=1}^{N_q} \left[\omega_q \exp(\sqrt{2}\sigma y_q)^{d_j} \exp\left(-\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) \exp(\sqrt{2}\sigma y_q)\right) \right] \right) - \quad (79)$$

$$\left. - \log(A_j(\boldsymbol{\theta})) + \log f(x_{ij,1,mis}|w_{ij}, \boldsymbol{\psi}) dx_{ij,1,mis} \right] \quad (80)$$

and the E-step:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \sum_{j=1}^J \left[\sum_{i=1}^{n_j} \sum_{m=1}^M (\delta_{ij} \log h(t_{ij}|\mathbf{x}_{ij}^{(m)})) + \right. \quad (81)$$

$$+ \log \left(\frac{1}{\sqrt{\pi}} \sum_{q=1}^{N_q} \left[\omega_q \exp(\sqrt{2}\sigma y_q)^{d_j} \exp\left(-\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}^{(m)}) \exp(\sqrt{2}\sigma y_q)\right) \right] \right) - \quad (82)$$

$$\left. - \log(A_j(\boldsymbol{\theta})) + \log f(x_{ij,1,mis}^{(m)}|w_{ij}, \boldsymbol{\psi}) \right] \quad (83)$$

8.2. Considering the Family Correlations

Given that family j has some subjects containing the missing PRS due to the sampling cost (maybe, need to confirm the previous paper), that not all subjects are being sampled for the PRS calculation. Since subjects within one family are correlated in some genetic associations, we intend to sample the missing PRS using a multivariate normal distribution while accounting for the kinship matrix. Using the same setting for the modelling of the PRS, the only difference is the variance-covariance matrix. Here, $\mathbf{X}_{j,1}|\mathbf{W}, \boldsymbol{\psi} \sim MVN(\mathbf{W}\boldsymbol{\psi} + \mathbf{u}, \sigma^2\mathbf{I})$. But, $\mathbf{u} \sim MVN(0, 2\sigma_u^2\mathbf{K})$ where \mathbf{K} is a $\mathbb{R}^{n \times n}$ kinship matrix with a diagonal of 0.5. The variance can be computed via $\text{Var}(X_{j,1}) = 2\sigma_u^2\mathbf{K} + \sigma^2\mathbf{I}$ when both the genetic variation and the residual errors contribute to the variance-covariance structure. Therefore, the marginal distribution of the PRS after integrating out the random effects will be $\mathbf{X}_{j,1} \sim MVN(\mathbf{W}\boldsymbol{\psi}, 2\sigma_u^2\mathbf{K} + \sigma^2\mathbf{I})$, and will be the distribution used for the Monte Carlo sampling.

M Step, Nelder-Mead, why? Gradient free? Variance?

9. Multiple Imputation with Monte Carlo Sampling

9.1. Missing Mechanisms

...A discussion on MCAR, MAR, and MNAR...

9.2. Multiple Imputations *(Be sure to report a conceptual statistical definition in the thesis)*

In the general framework, denote $\mathbf{X}_{j,1,mis}$ as the element of $\mathbf{X}_{j,1}$ when it's missing. Denote $\mathbf{X}_{j,1,obs}$ on the observed element. Adapting the idea from the MCEM, we can draw $\hat{\mathbf{X}}_{j,1,mis}$ from the same assumption of the distribution discussed in the section 8 for M times. But, the missing indicator \mathbf{R} should be addressed to attest the missing mechanism. The conditional distribution can be written as $f(\mathbf{X}_{j,1,mis}|\mathbf{X}_{j,1,obs}, \mathbf{W}, \mathbf{R}_j)$. Under the assumption of the MAR *(Confirm the paper)*,

$$f(\mathbf{X}_{j,1,mis}|\mathbf{X}_{j,1,obs}, \mathbf{W}, \mathbf{R}_j) = f(\mathbf{X}_{j,1,mis}|\mathbf{X}_{j,1,obs}, \mathbf{W}) \quad (84)$$

This results a series of M complete datasets containing all values observed with the fill of the Monte Carlo samples. Then with these M datasets, we run the analysis individually. Thus, the estimate of $\boldsymbol{\theta}$ can be calculated via

$$\hat{\boldsymbol{\theta}}_{MI} = \frac{1}{M} \sum_{m=1}^M \hat{\boldsymbol{\theta}}_m \quad (85)$$

Denote V the variance for m -th complete data inference variance, we first calculate the average variance of these analyses to be the within-imputation variance,

$$\hat{\mathbf{W}} = \frac{1}{M} \sum_{m=1}^M \hat{\mathbf{V}}_m \quad (86)$$

Under the unbiased variance estimator, the between-imputation variance is

$$\hat{\mathbf{B}} = \frac{1}{M-1} \sum_{m=1}^M (\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{MI})(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{MI})^\top \quad (87)$$

Combining these *(How? Are there any statistical/mathematical proof of this)*, we obtain the estimate of the variance of $\hat{\boldsymbol{\theta}}_{MI}$,

$$\hat{\mathbf{V}}_{MI} = \hat{\mathbf{W}} + (1 + \frac{1}{M})\hat{\mathbf{B}} \quad (88)$$

Discussions on why MI over MCEM or vise versa after the analysis...

10. Monte Carlo EM *(This section needs to be re-written to a definitional MCEM framework, maybe consult papers)*

The complete data log-likelihood for family j is $\ell_j(\boldsymbol{\theta}; h_{ij})$ where $\boldsymbol{\theta}$ consists all baseline parameters, and model coefficients β 's, as well as the frailty parameter ϕ . The E-step for complete data is:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \int \ell(\boldsymbol{\theta}; h_{ij}) \cdot f(x_{mis,i}|x_{obs,i}, z, \boldsymbol{\theta}^{(r)}, t_{ij}, \delta_{ij}, p_j) dx_{mis,ij} \quad (89)$$

We sample the size m_i for each i -th observation, $x_{i1}^*, \dots, x_{im_i}^*$ from the distribution $f(x_{mis,ij}|\cdot)$, and take $M = 1, \dots, m_i$, such that each X_{iM}^* depends on the iteration number for $r + 1$ iterations. In general:

$$\hat{Q}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \frac{1}{m_i} \sum_{M=1}^{m_i} \ell(x_{iM}^*, x_{obs,ij}, t_{ij}, \boldsymbol{\theta}, z_j) \quad (90)$$

More specifically,

1. We first initialize $m, \theta^{(0)}$, and start the burn-in.
2. Also, we set importance weights $w_t = 1$ for all $t = 1, \dots, m$.
3. At the burn-in iteration s , we generate $x_{miss,1}, \dots, x_{miss,m} \sim N(\mu_X | X_{obs}, \theta^{(s)}, z)$ using MCMC sample.
4. In the E-step, we estimate $Q(\theta | \theta^{(s)})$ by using the importance weights:

$$Q_m(\theta | \hat{\theta}^{(s)}) = \frac{\sum_{t=1}^m w_t \log f(X_{obs}, X_{miss,t} | \theta)}{\sum_{t=1}^m w_t} \quad (91)$$

5. Note the numerator is actually a weighted log-likelihood. In the M-step, we maximize $Q_m(\theta | \hat{\theta}^{(s)})$ to obtain $\hat{\theta}^{(s+1)}$.
6. Repeat (3.) - (5.) for s burn-in iterations.
7. Then re-initialize $\hat{\theta}^{(0)} = \hat{\theta}^{(s)}$
8. We generate $x_{miss,1}, \dots, x_{miss,m} \sim N(\mu_X | X_{obs}, \hat{\theta}^{(0)}, z)$ using MCMC sampler. At iteration $r + 1$
9. Compute the importance weights from the ratio of likelihood

$$w_t = \frac{L(\hat{\theta}^{(r)} | X_{miss,t}, X_{obs})}{L(\hat{\theta}^{(0)} | X_{miss,t}, X_{obs})} \quad (92)$$

10. Thus, the E-step can be written as

$$Q_m(\theta | \hat{\theta}^{(r)}) = \frac{\sum_{t=1}^m w_t \log f(X_{miss,t}, X_{obs} | \theta)}{\sum_{t=1}^m w_t} \quad (93)$$

11. Then M-step: we maximize $Q_m(\theta | \hat{\theta}^{(r)})$ to obtain $\hat{\theta}^{(r+1)}$.

This automated MCEM firstly optimizes the importance weights at burn-ins, then performs the actual EM to find $\hat{\theta}$. This importance weight ensures the imputation step of the missing data actually yields to the real distribution.

11. Correlated Frailty using Kinship Matrix

Family members are correlated within one family, that we denote K as the kinship correlation matrix among all observations. This matrix ensures those individuals not from the same family automatically have a correlation of 0. The likelihood construction needs multivariate form. For $\mathbf{Z} \sim \text{MVN}(0, \sigma^2 K)$, that K has the diagonal of 1. The likelihood is

$$L(\cdot) = \int_{\mathbb{R}^n} \prod_{i=1}^n (h(t|\mathbf{x}_i, \mathbf{z}_i))^{\delta_i} \exp(-H(t|\mathbf{x}_i, \mathbf{z}_i)) f(\mathbf{z}) d\mathbf{z} \quad (94)$$

$$= \int_{\mathbb{R}^n} \prod_{i=1}^n (h(t|\mathbf{x}_i))^{\delta_i} \exp(\mathbf{z}_i)^{\delta_i} \exp(-H(t|\mathbf{x}_i) \exp(\mathbf{z}_i)) f(\mathbf{z}) d\mathbf{z} \quad (95)$$

$$= \prod_{i=1}^n (h(t|\mathbf{x}_i))^{\delta_i} \int_{\mathbb{R}^n} \exp(\delta_i \mathbf{z}_i - H(t|\mathbf{x}_i) \exp(\mathbf{z}_i)) f(\mathbf{z}) d\mathbf{z} \quad (96)$$

177 Applying the Laplace approximation, and taking the log for the likelihood, we obtain

$$\ell(\cdot) = \sum_{i=1}^n \left[\delta_i \log h(t|\mathbf{x}_i) \right] + \sum_{i=1}^n \left[\delta_i \hat{\mathbf{z}} - H(t|\mathbf{x}_i) \exp(\hat{\mathbf{z}}) \right] - \frac{1}{2} \hat{\mathbf{z}}^\top \Sigma^{-1} \hat{\mathbf{z}} \quad (97)$$

178 such that $\Sigma = \sigma^2 K$. Also, we treat the random effect \mathbf{z} as a vector of parameters, and use
 179 outer-loop to search for the σ , and use inner-loop to search for other parameters (baseline
 180 parameters, and β) including \mathbf{z} . The process can be achieved via Newton-Raphson algorithm.
 181 For computational efficiency, we can set $\Sigma^{-1} = L^\top L$ through Cholesky Decomposition. In
 182 this way, $\mathbf{z}L \sim MVN(0, \sigma^2 I)$. In order to apply NR-algorithm, the gradient and the hessian
 183 are required. The gradient for parameters is:

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n \delta_i \mathbf{x}_i + \sum_{i=1}^n -H(t_i|\mathbf{x}_i) \mathbf{x}_i \exp(\mathbf{z}) \quad (98)$$

$$\frac{\partial \ell}{\partial \mathbf{z}} = \sum_{i=1}^n \delta_i - (t_i|\mathbf{x}_i) \exp(\hat{\mathbf{z}}) - \Sigma^{-1} \hat{\mathbf{z}} \quad (99)$$

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^n \frac{\delta_i}{\alpha} + \sum_{i=1}^n -\frac{H(t_i|\mathbf{x}_i) \exp(\hat{\mathbf{z}})}{\alpha} \quad (100)$$

$$\frac{\partial \ell}{\partial \lambda} = \sum_{i=1}^n \delta_i \left(\frac{1}{\lambda} + \log(t_i) \right) + \sum_{i=1}^n -H(t_i|\mathbf{x}_i) \exp(\hat{\mathbf{z}}) \log(t_i) \quad (101)$$

184 The hessian matrix element, i.e., second partial derivative is

$$\frac{\partial^2 \ell}{\partial \beta^\top \partial \beta} = \sum_{i=1}^n -H(t_i|\mathbf{x}_i) \exp(\hat{\mathbf{z}}) x_{ij} x_{ik} \quad (102)$$

$$\frac{\partial^2 \ell}{\partial \mathbf{z}^\top \partial \mathbf{z}} = \sum_{i=1}^n -H(t_i|\mathbf{x}_i) \exp(\hat{\mathbf{z}}) - \Sigma^{-1} \quad (103)$$

$$\frac{\partial^2 \ell}{\partial \alpha^2} = \sum_{i=1}^n -\frac{\delta_i}{\alpha^2} \quad (104)$$

$$\frac{\partial^2 \ell}{\partial \lambda^2} = \sum_{i=1}^n -\frac{\delta_i}{\lambda^2} - H(t_i | \mathbf{x}_i) \exp(\hat{\mathbf{z}}) \log(t_i)^2 \quad (105)$$

11.1. Proof of $\Sigma = LL^\top$

Every symmetric positive definite matrix Σ can be decomposed into $\Sigma = LL^\top$, where L is a lower triangular matrix with real and positive diagonal entries.

Proof. Set-ups:

1. Covariance matrix Σ is by definition symmetric and positive definite, e.g.

$$\Sigma = \begin{pmatrix} \sigma_{X_1}^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_{X_2}^2 \end{pmatrix} \quad (106)$$

such that $\mathbf{X}\Sigma\mathbf{X}^\top > 0$ always, and this matrix is symmetric.

2. Suppose \mathbf{X} has n observations, then Σ is $n \times n$, the first element is $\sigma_{11} > 0$ by definition (For simplicity, we use σ_{11} rather than it's square to denote the variance). Define $l_{11} = \sqrt{\sigma_{11}}$, to be the first element of L . For the first column of L , let $l_{j1} = \frac{\sigma_{j1}}{l_{11}}$ for $j = 2, \dots$

Induction step: Assume we have first $k-1$ columns of L , consider k -th column

- For the diagonal element $l_{kk} = \sqrt{\sigma_{kk} - \sum_{j=1}^{k-1} l_{kj}^2}$
- For off-diagonals,

$$l_{ik} = \frac{\sigma_{ik} - \sum_{j=1}^{k-1} l_{ij}l_{kj}}{l_{kk}} \quad (107)$$

for $i = k+1, \dots, n$.

with the repetition for each column $k = 2, \dots, n$, the top-left $k \times k$ submatrix of LL^\top matches that of Σ . For example, when $k = 3$,

$$\Sigma = \begin{pmatrix} \sigma_{11} & & \\ & \sigma_{22} & \\ & & \sigma_{33} \end{pmatrix} \quad (108)$$

and

$$L = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \quad (109)$$

then

$$LL^\top = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{pmatrix} \quad (110)$$

203 Take

$$\Sigma = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix} \quad (111)$$

204 Then by definition of Cholesky Decomposition, we can calculate $l_{11}^2 = \sigma_{11} \implies l_{11} = \sqrt{4} = 2$,
 205 and $l_{21} = \frac{\sigma_{21}}{l_{11}} = 2/2 = 1$, and $l_{31} = 1$. Similarly for l_{22}, l_{32}, l_{33} . Therefore,

$$L = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{2} & 0 \\ 1 & 0 & \sqrt{2} \end{pmatrix} \quad (112)$$

206 which implies

$$LL^\top = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{2} & 0 \\ 1 & 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix} = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix} = \Sigma \quad (113)$$

207

□

208 Essentially, the Cholesky Decomposition transforms the multivariate normal to a stan-
 209 dard multivariate normal. When $\mathbf{Z} \sim \mathcal{N}(0, \Sigma)$, let $\Sigma = \mathbf{L}\mathbf{L}^\top$, then $\mathbf{Y} = \mathbf{L}^{-1}\mathbf{Z} \sim \mathcal{N}(0, \mathbf{I})$
 210 that \mathbf{I} is the identity matrix, since $\mathbf{L}^{-1}\Sigma(\mathbf{L}^{-1})^\top = \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^\top(\mathbf{L}^{-1})^\top = \mathbf{I}$. This will simplify
 211 the computational process.

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.