Jiaqi's Thesis Progress Report (Updated Dec. 18)

Jiaqi Bi[a]

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

1. To Do List during the Christmas

1. Review the missing data slides
2. Find where went wrong in log-Normal Frailty
3. Complete the code for Correlated Frailty
4. Alternate the MCEM for Gamma Frailty case to use conditional sampling on the missing PRS (conditional on observed PRS, t, time-to-BC, family structure, $\boldsymbol{\theta}$)
5. Summarize the code within an R package to make sure it's readable
6. Start initializing the thesis
7. Next Meeting on Jan. 11, 2024
8. Add citations to this document!

2. Weibull Parametric Approach and MCEM Method

From the beginning of the discussion, I have obtained the model, i.e., the hazard function is

$$h_{ij}(t_{ij}|z_j) = h_0(t_{ij}) \exp(\beta_1 g_{ij} + \beta_2 x_{ij}) z_j \tag{1}$$

where g_{ij} is the genotype, or say mutation gene status for individual i in family j, and x_{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, or log-normal, to ensure the support is always non-negative. The Weibull baseline hazard function is defined as

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

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

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

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

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

$$H(t_{ij}|x_{ij}, g_{ij}, z_j) = \int_0^t h_{ij}(u|\cdot)du \tag{4}$$

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

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

and the survival function

$$S(t_{ij}|x_{ij}, g_{ij}, z_j) = \exp(-H(t_{ij}|\cdot)) = \exp(-\alpha\xi_{ij}z_j t_{ij}^{\lambda}) \tag{7}$$

Therefore, the likelihood can be written as

$$L(\beta_1, \beta_2, \lambda, \alpha; x_{ij}, g_{ij}, t_{ij}, \delta_{ij}, z_j) = \prod_{j=1}^{k} \int_0^{\infty} \prod_{i=1}^{n} (\alpha\lambda t_{ij}^{\lambda-1}\xi_{ij}z_j)^{\delta_{ij}} \exp(-\alpha\xi_{ij}z_j t_{ij}^{\lambda}) f(z)dz \tag{8}$$

So the log-likelihood is

$$\ell(\beta_1, \beta_2, \lambda, \alpha; x_{ij}, g_{ij}, t_{ij}, \delta_{ij}, z_j) = \sum_{j=1}^{k} \log \left[\int_0^{\infty} \prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}, z_j))^{\delta_{ij}} \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) f(z_j)dz_j \right] \tag{9}$$

3. Gamma Frailty

The Laplace transform of the frailty $f(z_j) = \text{Gamma}(v_j, v_j)$, for the simplicity of the mathematical expression, the following Laplace transform will ignore the subscript, denote $\mathscr{L}(f(z)) = \phi(\cdot)$:

$$\mathscr{L}(f(z)) = \phi(s) = \int_0^{\infty} e^{-sz} f(z)dz \tag{10}$$

$$= \int_0^{\infty} e^{-sz} \frac{v^v}{\Gamma(v)} z^{v-1} e^{-vz} dz \tag{11}$$

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

$$\phi(s) = \frac{v^v}{\Gamma(v)} \int_0^{\infty} e^{-(s+v)z} z^{v-1} dz = \frac{v^v}{\Gamma(v)} \cdot \frac{\Gamma(v)}{(s+v)^v} = (1 + \frac{s}{v})^{-v} \tag{12}$$

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

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

$$\phi(s)^{(d)} = (-1)^d \int_0^\infty z^d e^{-sz} f(z) dz \tag{13}$$

$$= (-1)^d \frac{(v+d-1)!}{(v-1)!(s+v)^d} (1+\frac{s}{v})^{-v} \tag{14}$$

for some function s that does not involve with z. Let $\boldsymbol{\theta} = (\beta_1, \beta_2, \alpha, \lambda)$, the log-likelihood is then written as

$$\ell(\boldsymbol{\theta}) = \sum_{j=1}^{k} \log \left[\int_0^\infty \prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}, z_j))^{\delta_{ij}} \exp(-H(t_{ij}|\mathbf{x}_{ij}, z_j)) f(z_j) dz_j \right] \tag{15}$$

$$= \sum_{j=1}^{k} \log \left[\int_0^\infty \prod_{i=1}^{n_j} (z_j h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \exp(-z_j H(t_{ij}|\mathbf{x}_{ij})) f(z_j) dz_j \right] \tag{16}$$

$$= \sum_{j=1}^{k} \log \left[\prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \int_0^\infty z_j^{d_j} \exp(-z_j \sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij})) f(z_j) dz_j \right] \tag{17}$$

$$= \sum_{j=1}^{k} \log \left[\prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \frac{(v+d_j-1)!}{(v-1)!(\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) + v)^{d_j}} \Big(1 + \frac{\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij})}{v}\Big)^{-v} \right] \tag{18}$$

$$= \sum_{j=1}^{k} \log \left[\prod_{i=1}^{n_j} ((h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}}) \frac{(v+d_j-1)!}{v! v^{d_j-1}} (1 + \frac{\sum_{i=1}^{n_j} (\alpha t_{ij}^{\lambda} \xi_{ij})}{v})^{-v-d_j} \right] \tag{19}$$

$$= \sum_{j=1}^{k} \log \left[(h(\cdot))^{\delta_{ij}} \frac{(v+d_j-1)!}{v! v^{d_j-1}} (1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}))}{v})^{-v-d_j} \right] \tag{20}$$

$$= \sum_{j=1}^{k} \left[\sum_i (\delta_{ij} \log h(\cdot)) + \log \Big(\frac{(v+d_j-1)!}{v! v^{d_j-1}} (1 + \frac{\sum_{i=1}^{n_j} (H(t_{ij}|\mathbf{x}_{ij}))}{v})^{-v-d_j} \Big) \right] \tag{21}$$

For each family j, the ascertainment is defined to be the probability of the proband p being ascertained by the age a_{j_p} at examination, denoting A_j. Applying the ascertainment correction for the log-likelihood in family j:

$$\ell(\cdot) = \ell_j(\cdot) - \log A_j(\cdot) \tag{22}$$

note we can still apply Laplace transform here, such that

$$A_j(\cdot) = 1 - S_{j_p}(a_{j_p}|X_{j_p}) \tag{23}$$
$$= 1 - \int_Z S_{j_p}(a_{j_p}|X_{j_p}, z_j) f(z_j) dz_j \tag{24}$$
$$= 1 - \int_Z \exp(-z_j \cdot H_{j_p}(a_{j_p}|X_{j_p})) f(z_j) dz_j \tag{25}$$
$$= 1 - (1 + \frac{H_{j_p}(a_{j_p}|X_{j_p})}{k})^{-k} \tag{26}$$

4. 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

$$E(\log Z) = 0 \tag{27}$$
$$\text{Var}(\log Z) = \sigma^2 \tag{28}$$

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

$$f(z_j) = \frac{1}{\sqrt{2\pi}\sigma} z_j^{-1} \exp(-\frac{\log(z_j)^2}{2\sigma^2}) \tag{29}$$

The Laplace transform is then

$$\phi(s) = \mathscr{L}(f_Z)(s) = \int_0^\infty \exp(-sz) \cdot f(z) dz \tag{30}$$

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:

$$\phi(s)^d = \int_{-\infty}^\infty z^d \exp(-sz) \cdot \frac{1}{\exp(\sqrt{2}\sigma y)\sigma\sqrt{2\pi}} \cdot \exp(-y^2) \cdot \sqrt{2}\sigma \exp(\sqrt{2}\sigma y) dy \tag{31}$$
$$= \int_{-\infty}^\infty \exp(\sqrt{2}\sigma y)^d \exp(-s \exp(\sqrt{2}\sigma y)) \cdot \frac{1}{\sqrt{\pi}} \exp(-y^2) dy \tag{32}$$

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:*

$$\int_{-\infty}^\infty \exp(-x^2) f(x) dx \approx \sum_{i=1}^n \omega_i f(x_i) \tag{33}$$

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*

$$\omega_i = \frac{2^{n-1} n! \sqrt{n}}{n^2 [H_{n-1}(x_i)]^2} \tag{34}$$

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

$$\phi(s)^d = \frac{1}{\sqrt{\pi}} \sum_{p=1}^{N_p} \omega_p \exp(-s \exp(\sqrt{2}\sigma y_p)) \exp(\sqrt{2}\sigma y_p)^d \tag{35}$$

Thus, substituting into the log-likelihood:

$$\ell_j(\cdot) = \sum_{i=1}^{n_j} \delta_{ij} \log(h(t_{ij}|\mathbf{x}_{ij})) + \log\Big(\frac{1}{\sqrt{\pi}} \sum_{p=1}^{N_p} \Big[\omega_p \exp(\sqrt{2}\sigma y_p)^{d_{ij}} \exp\Big(-\sum_{i=1}^{n_j} H(t_{ij}|\mathbf{x}_{ij}) \exp(\sqrt{2}\sigma y_p)\Big)\Big]\Big) \tag{36}$$

5. Incorporating the Kinship Matrix to Log-Normal Frailty

Family members are correlated within one family, that we denote K as the kinship correlation matrix among all individuals i. Those who do not share familial relationships automatically have correlation of 0. The kinship matrix K has a diagonal of 0.5. The likelihood construction needs multivariate form. For $\mathbf{Z} \sim \text{MVN}(0, \sigma^2 K)$, the log-likelihood for family j is

$$\ell_j(\cdot) = \log\left[\prod_{i=1}^{n_j} (h(t_{ij}|\mathbf{x}_{ij}))^{\delta_{ij}} \int_{\mathbb{R}^{n_j}} \prod_{i=1}^{n_j} z_{ij}^{\delta_{ij}} \exp(\sum_{i=1}^{n_j} z_{ij} H(t_{ij}|\mathbf{x}_{ij})) f(\mathbf{z}_j) d\mathbf{z}_j\right] \tag{37}$$

Applying the Cholesky Decomposition, the integrand is

$$\int_{\mathbb{R}^{n_j}} \prod_{i=1}^{n_j} (\mathbf{LZ})_{ij}^{d_j} \exp(-\sum_{i=1}^{n_j} (\mathbf{LZ})_{ij} H(t_{ij}|\mathbf{x}_{ij})) \Phi(\mathbf{z}) d\mathbf{z} \tag{38}$$

Applying the Gauss Hermite, it becomes

$$\sum_{p=1}^{N_p} \omega_p (\mathbf{LZ}_p)_{ij}^{d_j} \exp(-\sum_{i=1}^{n_j} (\mathbf{LZ}_p)_{ij} H(t_{ij}|\mathbf{x}_{ij})) \tag{39}$$

such that $\mathbf{L}$ is the Cholesky Decomposition, which $\Sigma = \mathbf{L}\mathbf{L}^\top$. And ω_p is the multivariate Gauss Hermite weights, $\mathbf{Z}_p$ is the nodes. The Cholesky Decomposition works well for positive definite matrices, it simplifies the computation of matrix operations, especially in solving systems of linear equations and inverting matrices. In our case, the covariance matrix Σ is symmetric and positive definite, which the Cholesky Decomposition is ideally suited.

5.1. *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 $\mathbf{\Sigma}$ is by definition symmetric and positive definite, e.g.

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{X_1}^2 & Cov(X_1, X_2) \\ Cov(X_1, X_2) & \sigma_{X_2}^2 \end{pmatrix} \tag{40}$$

   such that $\mathbf{X\Sigma X}^\top > 0$ 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 $j = 2, ...$

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,

$$l_{ik} = \frac{\sigma_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj}}{l_{kk}} \tag{41}$$

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

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

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_{11} & & \\ & \sigma_{22} & \\ & & \sigma_{33} \end{pmatrix} \tag{42}$$

and

$$L = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \tag{43}$$

then

$$LL^\top = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{pmatrix} \tag{44}$$

Take

$$\Sigma = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix} \tag{45}$$

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

$$L = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{2} & 0 \\ 1 & 0 & \sqrt{2} \end{pmatrix} \tag{46}$$

which implies

$$LL^\top = \begin{pmatrix} 2 & 0 & 0 \\ 1 & \sqrt{2} & 0 \\ 1 & 0 & \sqrt{2} \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix} = \begin{pmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix} = \Sigma \tag{47}$$

□

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

6. Monte Carlo EM

WLOG, the complete data log-likelihood for family j is $\ell(\boldsymbol{\theta}; h_{ij})$ where $\boldsymbol{\theta}$ consists all baseline parameters, and model coefficients say β. Suppose PRS and genotype are correlated somehow, denote ρ be their correlation. For each cluster j, the E-step is (Scenario 1: PRS observed but genotype missing):

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \int \ell(\boldsymbol{\theta}; h_{ij}) \cdot f(g_{mis,i}|h_{ij}(t_{ij}|g_{obs,ij}, x_{ij}, z_j), \rho, \boldsymbol{\theta}^{(r)}) dg_{mis,ij} \tag{48}$$

In the scenario 2 that PRS missing while genotype observed:

$$Q_i(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = \int \ell(\boldsymbol{\theta}; h_{ij}) \cdot f(x_{mis,i}|h_{ij}(t_{ij}|x_{obs,ij}, g_{ij}, z_j), \rho, \boldsymbol{\theta}^{(r)}) dx_{mis,ij} \tag{49}$$

In the scenario 3 that both are missing, considering the joint distribution of these two covariates:

$$Q_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \int\int_{(X,G)} \ell(\cdot) f(x_{mis,ij}, g_{mis,ij}|h_{ij}(t_{ij}|x_{obs,ij}, g_{ij}, z_j), \boldsymbol{\theta}^{(r)}) dx_{mis,ij} dg_{mis,ij} \tag{50}$$

where $(X, G) \sim f(x, g|\cdot)$, and can be obtained through Gibbs Sampling, sample the size m_i for each i-th observation, $x_{i1}^*, ..., x_{im_i}^*$ from the distribution $f(x_{mis,ij}|\cdot)$, and take $M = 1, ..., m_i$, such that each X_{iM}^* depends on the iteration number for $r+1$ iterations. In general:

$$\hat{Q}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(r)}) = \frac{1}{m_i} \sum_{M=1}^{m_i} \ell(x_{iM}^*, x_{obs,ij}, t_{ij}, \boldsymbol{\theta}, z_j) \tag{51}$$

7. Stochastic EM?

8. Multiple Imputation Method

This method is easier to implement, because MICE package already incorporates fully conditional specification where each missing variable is imputed one at a time, conditional on others in the dataset. However, I only need to write a self-defined function to include the random effect in the imputation-step. All methods provided by MICE fail to include the frailty scenario in survival analysis. There are many sources that they considered random effects in the linear mixed effect model. I believe survival analysis with frailty should be similar.

```
### MICE for coxph with frailty
library(mice)
brca1_prs_mice$na <- nelsonaalen(data = brca1_prs_mice, timevar = "timeBC",
                                 statusvar = "BC")
micesurv0 <- mice(brca1_prs_mice, maxit = 0, method = "rf")
micesurvmethod <- micesurv0$method
micesurvpred <- micesurv0$predictorMatrix
micesurvmethod[c("PRS")] <- "rf"
micesurvmethod[c("mgene")] <- "rf"
micesurvpred[, "indID"] <- 0

micesurv <- mice(brca1_prs_mice, method = micesurvmethod, predictorMatrix = micesurvpred,
                 m = 5, seed = 123)
results_mice <- with(micesurv, coxph(Surv(timeBC, BC) ~ mgene + PRS + frailty(famID, distrib
summary(pool(results_mice), exponentiate = TRUE) # Here the MI does not incorporate random e
#
```

Figure 1: MICE Code

Here I was trying to use random forest methods to impute both PRS and genotype, the performance of PRS seems good because the coefficient is close to the complete case analysis, but there is an inflation on the coefficient on genotype.