FRAILTY MODEL WITH SPLINE ESTIMATED NONPARAMETRIC HAZARD FUNCTION

Pang Du and Shuangge Ma

Virginia Tech and Yale University

Supplementary Material

S1. Derivation of Cross-validation Score

In this section, we shall derive the cross-validation score (7) from the Kullback-Leibler distance (6). Dropping the terms that do not involve $\Theta_\Lambda = (\eta_\Lambda, \mathbf{b}_\Lambda)$, one can estimate the remaining part of (6) by

$$\frac{1}{n}\sum_{i=1}^{n}\int_{Z_i}^{X_i} e^{\eta_\Lambda(t,U_i)+\mathbf{z}_i^T\mathbf{b}_\Lambda}dt - \frac{1}{n}\sum_{i=1}^{n}\int_{Z_i}^{X_i}[\eta_\Lambda(t,U_i)+\mathbf{z}_i^T\mathbf{b}_\Lambda]e^{\eta(t,U_i)+\mathbf{z}_i^T\mathbf{b}}dt. \qquad \text{(S1)}$$

The first term of (S1) is available through Θ_Λ. But the second term, denoted by $\mu_\Theta(\Theta_\Lambda)$, involves the unknown Θ and has to be estimated.

The unknown Θ appears in the exponential power within $\mu_\Theta(\Theta_\Lambda)$, which is not friendly to deal with. So our first step is to transform $\mu_\Theta(\Theta_\Lambda)$ using the counting processes for frailty models introduced by Nielsen, Gill, Andersen, and Sørensen (1992).

Let $N(t) = I_{[X\le t,\delta=1]}$ be the event process and $Y(t) = I_{[Z<t\le X]}$ the at-risk process. Write $A(t) = \int_0^t Y(s)E_{\mathbf{b}}[e^{\eta(s,U)+\mathbf{z}^T\mathbf{b}}]ds$, with $E_{\mathbf{b}}$ being the expectation with respect to $\mathbf{b}$. Define $M(t) = N(t) - A(t)$. Then conditioning on Z and U, $M(t)$ is a martingale. By the martingale transform theorem, the integral $\int_0^t f(s,U)dM(s)$, given Z and U, is also a martingale for $f(t,u)$ independent of X and continuous in t, $\forall u \in \mathcal{U}$. Thus, one has $E[\int_{\mathcal{T}} f(t,U)dM(t)] = 0$, where the expectation is with respect to $\mathbf{b}$, Z, X and U. "Estimating zero" by the corresponding sample mean, one has

$$0 \approx \frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i f(X_i,U_i) - \int_{Z_i}^{X_i} f(t,U_i)e^{\eta(t,U_i)+\mathbf{z}_i^T\mathbf{b}}dt\right\}. \qquad \text{(S2)}$$

A naive estimate of $\mu_\Theta(\Theta_\Lambda)$ can be obtained by setting $f(t,U_i) = \eta_\Lambda(t,U_i) + \mathbf{z}_i^T\mathbf{b}_\Lambda$:

$$\tilde{\mu}_\Theta(\Theta_\Lambda) = \frac{1}{n}\sum_{i=1}^{n}\int_{Z_i}^{X_i}(\eta_\Lambda(t,U_i)+\mathbf{z}_i^T\mathbf{b}_\Lambda)e^{\eta(t,U_i)+\mathbf{z}_i^T\mathbf{b}}dt \approx \frac{1}{n}\sum_{i=1}^{n}\delta_i(\eta_\Lambda(X_i,U_i)+\mathbf{z}_i^T\mathbf{b}_\Lambda) \qquad \text{(S3)}$$

However, the resulting estimate of the Kullback-Leibler distance would simply be the negative log likelihood, clearly favoring $\lambda = 0$. The naive estimate (S3) is biased since the samples (X_i, U_i) contribute to the estimate Θ_Λ. An alternative to Θ_Λ is its approximation through standard cross-validation. Consider the following delete-one version of a quadratic approximation to (1) at $\tilde{\Theta}$,

$$-\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^{n}\delta_j(\eta(X_j,U_j)+\mathbf{z}_j^T\mathbf{b})+\mu_{\tilde{\Theta}}(\Theta)-V_{\tilde{\Theta}}(\tilde{\Theta},\Theta)+\frac{1}{2}V_{\tilde{\Theta}}(\Theta,\Theta)+\frac{1}{2n}\mathbf{b}^T\Sigma\mathbf{b}+\frac{\lambda}{2}J(\eta), \quad \text{(S4)}$$

where $V_{\tilde{\Theta}}(\tilde{\Theta},\Theta) = V_{\tilde{\Theta}}(\tilde{\eta}+\mathbf{z}^T\tilde{\mathbf{b}}, \eta+\mathbf{z}^T\mathbf{b})$, and $V_{\tilde{\Theta}}(\Theta,\Theta) = V_{\tilde{\Theta}}(\eta+\mathbf{z}^T\mathbf{b}, \eta+\mathbf{z}^T\mathbf{b})$, with $V_{\tilde{\Theta}}$ defined in Section 2.2.

Set $\tilde{\Theta} = \Theta_\Lambda$ in (S4). Denote the resulting minimizer by $\Theta^{[i]}_{\Lambda,\tilde{\Theta}}$. Let $\mathbf{a} = (\mathbf{b}^T, \mathbf{c}^T, \mathbf{d}^T)^T$. Then a straightforward calculation using the delete-one version of (5) for (S4) yields the coefficient for $\Theta^{[i]}_{\Lambda,\tilde{\Theta}}$:

$$\mathbf{a}^{[i]}_\Lambda = \mathbf{a}_\Lambda + \frac{H^{-1}K\mathbf{1}}{n(n-1)} - \delta_i\frac{H^{-1}\boldsymbol{\psi}(X_i,U_i)}{n-1},$$

where H is the hessian matrix on the left hand side of (5), K is a $(p+q+m)\times n$ matrix with columns $\boldsymbol{\psi}(X_i,U_i) = (\mathbf{z}_i^T, \boldsymbol{\xi}(X_i,U_i)^T, \boldsymbol{\phi}(X_i,U_i)^T)^T$, and $\mathbf{1}$ is a $(p+q+m)$-vector of all 1's. Hence

$$\begin{aligned}\eta^{[i]}_{\Lambda,\tilde{\Theta}}(X_i,U_i)+\mathbf{z}_i^T\mathbf{b}^{[i]}_{\Lambda,\tilde{\Theta}} &= \boldsymbol{\psi}(X_i,U_i)^T\mathbf{a}^{[i]}_\Lambda\\ &= [\eta_\Lambda(X_i,U_i)+\mathbf{z}_i^T\mathbf{b}_\Lambda] - \frac{1}{n-1}\boldsymbol{\psi}(X_i,U_i)^TH^{-1}(\delta_i\boldsymbol{\psi}(X_i,U_i)-K\mathbf{1}/n) \quad \text{(S5)}\end{aligned}$$

Substituting $\eta_\Lambda(X_i,U_i)+\mathbf{z}_i^T\mathbf{b}_\Lambda$ in (S3) by (S5) and then plugging the resulting estimate into (S1) yields the cross-validation score (7).

S2. Approximate Posterior Mean and Variance

In this part, we will illustrate the Bayes model for the penalized likelihood (1), and use it to derive the approximate posterior mean and variance of $\eta(x)+\mathbf{z}^T\mathbf{b}$ for any given $(x,\mathbf{z})$.

Since $\eta_0 = \boldsymbol{\phi}^T\mathbf{d}$ and $\eta_1 = \boldsymbol{\xi}^T\mathbf{c}$, the priors on η_0 and η_1 give a diffuse prior on $\mathbf{d}$ and a Gaussian prior $N(\mathbf{0}, Q^+/(n\lambda))$ on $\mathbf{c}$. The frailty $\mathbf{b}\sim N(0,\Sigma^{-1})$. Let $\mathbf{a} = (\mathbf{b}^T,\mathbf{c}^T,\mathbf{d}^T)^T$. The posterior likelihood of $\mathbf{a}$ given data $\mathbf{X} = \{(Z_i, X_i, \delta_i, U_i, \mathbf{z}_i), i = 1,\dots,n\}$ is proportional to the joint likelihood, which is of the form

$$\begin{aligned}p(\mathbf{X}|\mathbf{a})q(\mathbf{a}) \propto \exp\Bigg\{\sum_{i=1}^n\Big\{\delta_i(\boldsymbol{\phi}_i^T\mathbf{d}+\boldsymbol{\xi}_i^T\mathbf{c}+\mathbf{z}_i^T\mathbf{b}) - \int_{Z_i}^{X_i}\exp(\boldsymbol{\phi}(t,U_i)^T\mathbf{d}+\boldsymbol{\xi}(t,U_i)^T\mathbf{c}+\mathbf{z}_i^T\mathbf{b})dt\Big\}\\ -\frac{1}{2}\mathbf{b}^T\Sigma\mathbf{b}-\frac{n\lambda}{2}\mathbf{c}^TQ\mathbf{c}\Bigg\}, \quad \text{(S6)}\end{aligned}$$

with the exponent on the right hand side simply being the penalized likelihood (1) and (4) multiplied by n. Through a second-order Taylor expansion of the integral term at $\mathbf{a}_\Lambda$, the exponent in (S6) can be approximated by

$$-\frac{1}{2}(\mathbf{a}-\mathbf{a}_\Lambda)^T(nH)(\mathbf{a}-\mathbf{a}_\Lambda)+C, \tag{S7}$$

where H is the hessian matrix in (5) and C is a constant with respect to $\mathbf{b}$. Hence the approximate posterior distribution through (S7) is Gaussian with mean $\mathbf{a}_\Lambda$ and covariance H^{-1}/n. It follows that the approximate posterior mean of $\eta(x)+\mathbf{z}^T\mathbf{b}$ is $\eta_\Lambda(x)+\mathbf{z}^T\mathbf{b}_\Lambda=\boldsymbol{\psi}(x)^T\mathbf{a}_\Lambda$ and the approximate posterior variance is $\boldsymbol{\psi}^T(x)H^{-1}\boldsymbol{\psi}(x)/n$.

S3. Proof of Theorem 1

We will first prove identifiability of the model with any fixed p. Let $D=(Z,X,\delta,U,\mathbf{z})$ be a representative observation and $f(D;\mathbf{b},\eta)$ be the probability function of D with parameter value $(\mathbf{b},\eta)$. Then the proposed model is identifiable if and only if

$$\int(\sqrt{f(D;\mathbf{b},\eta)}-\sqrt{f(D;\mathbf{b}^*,\eta^*)})^2d\mu(D)=0 \tag{S8}$$

implies $(\mathbf{b},\eta)=(\mathbf{b}^*,\eta^*)$. We note that equation (S8) leads to

$$\delta(\eta+\mathbf{z}^T\mathbf{b})-\int_Z^X\exp(\eta+\mathbf{z}^T\mathbf{b})dt=\delta(\eta^*+\mathbf{z}^T\mathbf{b}^*)-\int_Z^X\exp(\eta^*+\mathbf{z}^T\mathbf{b}^*)dt.$$

Taking partial derivatives with respect to $\mathbf{z}$, Z, and X sequentially leads to

$$(e^{\eta^*(X,U)}-e^{\eta^*(Z,U)})e^{\mathbf{z}^T\mathbf{b}^*}\mathbf{b}^*=(e^{\eta(X,U)}-e^{\eta(Z,U)})e^{\mathbf{z}^T\mathbf{b}}\mathbf{b},$$

which implies $(e^{\eta^*(X,U)}-e^{\eta^*(Z,U)})=M_2(e^{\eta(X,U)}-e^{\eta(Z,U)})$ with a constant M_2. Simple calculations show that this leads to $\eta=\eta^*$ and $\mathbf{b}=\mathbf{b}^*$. In this above proof, we have further assumed that U is not a deterministic function of $\mathbf{z}$, which can be easily satisfied.

Next, let's look at the consistency of the estimates in the simple case when the covariance matrix $\Sigma=\sigma^{-2}I$. Assumption A4 indicates that $\sigma^{-2}=O_p(n^{\frac{1}{2s_0+1}})$. We have assumed that $(\mathbf{b}_T,\eta_T)$ is bounded. The first term in the likelihood function is linear, and the second term is exponential. So if a component of $(\hat{\mathbf{b}},\hat{\eta})$ is unbounded, then the second term of the likelihood, which is negative, will dominate. The penalty term is always negative. We can thus conclude that $(\hat{\mathbf{b}},\hat{\eta})$ is asymptotically bounded. We note that, in the above arguments, the boundedness of $\hat{\mathbf{b}}$ and $\mathbf{b}_T$ should be interpreted as component-wise.

Let $N_{[]}(\epsilon,\mathbb{G},||\cdot||)$ denote the bracketing number of ϵ-brackets covering a subset $\mathbb{G}$ of a real function space with norm $||\cdot||$. van de Geer (2000) shows that for the class $\mathbb{H}=\{h:[0,1]\to[0,1],\ \int(h^{(s_0)}(x))^2dx<1\}$, we have $\log N_{[]}(\epsilon,\mathbb{H},L_2)\le M_5\epsilon^{-\frac{1}{s_0}}$ for fixed $s_0\ge1$ and all ϵ, where M_5 is a fixed constant.

First, we note that, if p is bounded, then the following proof can be modified to establish that $\hat{\mathbf{b}}$ is $\sqrt{n}$ consistent for $\mathbf{b}_T$. Particularly, in what follows, we can first establish the $n^{s_0/(2s_0+1)}$ convergence rate for $\hat{\mathbf{b}}$ and $\hat{\eta}$. Then we can employ the general theorem presented in Huang

(1996) to establish the $\sqrt{n}$ consistency of $\hat{\mathbf{b}}$. With a diverging number of p, beyond assumptions A1-A5, we will also assume that $\frac{min_i P(z=i)}{max_i P(z=i)}$ is uniformly bounded away from 0, as $n \to \infty$. Intuitively, we assume that, as $n \to \infty$, the number of observations with different z values increases with asymptotically the same rate. This helps avoid the possible "partial consistency" situation, that is, as $n \to \infty$ the number of observations with certain z values remains small. Under this assumption, when $p \to \infty$ as $n \to \infty$, we can first carry out the proposed estimation with a fixed number of p. Denote the estimate so obtained as $\tilde{\mathbf{b}}$. Using results with fixed p, we can establish that component-wise $\tilde{\mathbf{b}} - \mathbf{b}_T \sim O(\sqrt{p/n})$, or $\tilde{\mathbf{b}} = \mathbf{b}_T \pm u\sqrt{p/n}$, where u is component-wise bounded. Thus, in the asymptotic study that follows, we are able to restrict our $\hat{\mathbf{b}}$ to be within $\tilde{\mathbf{b}} \pm u\sqrt{p/n}$. We note that, although we will localize our estimate, we do not assume we know the local region *a priori*. Rather, the local region is obtained via estimation.

Since $(\hat{\mathbf{b}}, \hat{\eta})$ minimizes (1), we have

$$\mathrm{P}_n l(\hat{\mathbf{b}}, \hat{\eta}) - \frac{1}{2n\sigma^2}\hat{\mathbf{b}}^T\hat{\mathbf{b}} - \frac{\lambda}{2}J(\hat{\eta}) \geq \mathrm{P}_n l(\mathbf{b}_T, \eta_T) - \frac{1}{2n\sigma^2}\mathbf{b}_T^T\mathbf{b}_T - \frac{\lambda}{2}J(\eta_T), \tag{S9}$$

which is equivalent to

$$\begin{aligned}\frac{1}{2n\sigma^2}\hat{\mathbf{b}}^T\hat{\mathbf{b}} + \frac{\lambda}{2}J(\hat{\eta}) + \mathrm{P}[l(\mathbf{b}_T, \eta_T) - l(\hat{\mathbf{b}}, \hat{\eta})]&\\ \leq \frac{1}{2n\sigma^2}\mathbf{b}_T^T\mathbf{b}_T + \frac{\lambda}{2}J(\eta_T) - (\mathrm{P}_n - \mathrm{P})[l(\mathbf{b}_T, \eta_T) - l(\hat{\mathbf{b}}, \hat{\eta})]&\end{aligned} \tag{S10}$$

Applying the entropy result for $l(\mathbf{b}, \eta)$, we have

$$(\mathrm{P}_n - \mathrm{P})[l(\mathbf{b}_T, \eta_T) - l(\hat{\mathbf{b}}, \hat{\eta})] = (1 + J^{1/2}(\hat{\eta}) + J^{1/2}(\eta_T))o_p(n^{-1/2} + ||\hat{\mathbf{b}} - \mathbf{b}_T||). \tag{S11}$$

Simple calculations after combining the above two inequalities shows $\lambda J(\hat{\eta}) = o_p(1)$. Equations (S10), (S11) and assumption A3 then imply that

$$d_\wedge^2 \leq o_p(1) + (1 + J^{1/2}(\hat{\eta}) + J^{1/2}(\eta_T))o_p(n^{-1/2} + ||\hat{\mathbf{b}} - \mathbf{b}_T||),$$

where $d_\wedge = d((\hat{\mathbf{b}}, \hat{\eta}), (\mathbf{b}_T, \eta_T))$. Consistency of the estimate thus holds.

To establish the rate of convergence, we use the following result.

Theorem (van de Geer 2000, Page 79). Consider a uniformly bounded class of functions Γ, with $\sup_{\gamma\in\Gamma}|\gamma - \gamma_0| < \infty$ with a fixed $\gamma_0 \in \Gamma$, and $\log N_{[]}(\epsilon, \Gamma, P) \leq M_7\epsilon^{-b}$ for all $\epsilon > 0$, where $b \in (0, 2)$ and M_7 a fixed constant. Then for $\delta_n = n^{-1/(2+b)}$, $\sup_{\gamma\in\Gamma} \frac{|(\mathrm{P}_n - \mathrm{P})(\gamma-\gamma_0)|}{||\gamma-\gamma_0||_2^{1-b/2} \vee \sqrt{n}\delta_n^2} = O_p(n^{-1/2})$, where $x \vee y = \max(x, y)$.

Combining assumption A3, (S10) and the above theorem with $b = \frac{1}{s_0}$ yields

$$\begin{aligned}&\frac{1}{2n\sigma^2}\hat{\mathbf{b}}^T\hat{\mathbf{b}} + \frac{\lambda}{2}J(\hat{\eta}) + M_3 d_\wedge^2 \leq \frac{1}{2n\sigma^2}\mathbf{b}_T^T\mathbf{b}_T + \frac{\lambda}{2}J(\eta_T)\\ &+ (1 + J^{1/2}(\hat{\eta}) + J^{1/2}(\eta_T)) \times O_p(n^{-1/2}) \times (||\hat{\eta} - \eta_T||^{1-\frac{1}{2s_0}} + ||\hat{\mathbf{b}} - \mathbf{b}_T||^{1-\frac{1}{2s_0}} + n^{\frac{1}{2}-\frac{2s_0}{2s_0+1}}).\end{aligned}$$

From the above equation, we can get that

$$
\begin{aligned}
&\frac{\lambda}{2}J(\hat{\eta}) \leq \frac{1}{2n\sigma^2}\mathbf{b}_T^T\mathbf{b}_T + \frac{\lambda}{2}J(\eta_T)\\
&+(1+J^{1/2}(\hat{\eta})+J^{1/2}(\eta_T)) \times O_p(n^{-1/2}) \times (||\hat{\eta}-\eta_T||^{1-\frac{1}{2s_0}} + ||\hat{\mathbf{b}}-\mathbf{b}_T||^{1-\frac{1}{2s_0}} + n^{\frac{1}{2}-\frac{2s_0}{2s_0+1}})\\
&M_3 d_\wedge^2 \leq \frac{1}{2n\sigma^2}\mathbf{b}_T^T\mathbf{b}_T + \frac{\lambda}{2}J(\eta_T)\\
&+(1+J^{1/2}(\hat{\eta})+J^{1/2}(\eta_T)) \times O_p(n^{-1/2}) \times (||\hat{\eta}-\eta_T||^{1-\frac{1}{2s_0}} + ||\hat{\mathbf{b}}-\mathbf{b}_T||^{1-\frac{1}{2s_0}} + n^{\frac{1}{2}-\frac{2s_0}{2s_0+1}}).
\end{aligned}
$$

Simple calculations yield that $J(\hat{\eta}) = O_p(1)$ and $d((\hat{\mathbf{b}},\hat{\eta}),(\mathbf{b}_T,\eta_T)) = d_\wedge = O_p(n^{-\frac{s_0}{2s_0+1}})$.

Note that the above proof of consistency and convergence rate relies only on the fact that the order of σ^{-2} is in control. For a general Σ, the proof can thus be easily modified using the eigendecomposition of Σ.