Survival Time Analysis with Mixtures of Accelerated Failure Time (AFT) Model Experts

Daniel Maturana and Abel Valdebenito

May 19, 2017

1 Accelerated Failure Time Models

A parametric model for survival time data.

Survival Function for AFT

$$S(t|x) = \psi\left(\frac{\log t - x'\beta}{\sigma}\right)$$

For log-normal AFT,

$$\psi(\cdot) = 1 - \Phi(\cdot)$$

So

$$\begin{aligned}
F(t) &= 1 - S(t) \\
&= \Phi\left(\frac{\log t - x'\beta}{\sigma}\right) \\
f(t) &= F'(t) \\
&= \phi\left(\frac{\log t - x'\beta}{\sigma}\right)\frac{1}{\sigma t} \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(\log t - x'\beta)^2}{2\sigma^2}\right)\frac{1}{\sigma t} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(\log t - x'\beta)^2}{2\sigma^2}\right)\frac{1}{t} \\
&= \mathcal{N}(\log t|x'\beta, \sigma^2)\frac{1}{t} \\
&= \mathcal{LN}(t|x'\beta, \sigma^2)
\end{aligned}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ is the normal density of mean μ and σ^2, and $\mathcal{LN}(\cdot|\mu, \sigma^2)$ is the corresponding log-normal density.

Also

$$\begin{aligned}
\log T &= x'\beta + \sigma\varepsilon \\
\varepsilon &\sim \mathcal{N}(0, 1)
\end{aligned}$$

Or

$$
\begin{aligned}
T &= \exp(x'\beta)v \\
v &= \exp(\sigma\varepsilon) \\
\varepsilon &\sim \mathcal{N}(0,1)
\end{aligned}
$$

If data is censored, likelihood may be calculated by integrating f over ranges where the true t may lie.

2 Mixtures of Experts of AFTs

Idea: Mixture of log-normal AFTs where weight of each mixture components depends on the covariates x; each AFT component becomes an "expert" for certain region of the input space.

Let $\theta = \left\{\beta_{1:J}, \rho_{1:J}, \sigma^2_{1:J}\right\}$.

$$
f(t|x,\theta) = \sum_{j}^{J} p_j(x, \rho_{1:J}) f_j(t|x, \beta_j, \sigma_j^2)
$$

Where f_j are log-normal AFT. We use a logistic link

$$
p_j(x, \rho_{1:J}) = \frac{\exp(x'\rho_j)}{\sum_l^J \exp(x'\rho_l)}
$$

Also

$$
F(t|x,\theta) = \sum_{j}^{J} p_j(x, \rho_{1:J}) F_j(t|x, \beta_j, \sigma_j^2)
$$

3 Likelihood

We asume that failture times are independently distributed and independent from the censoring process. For each individual we observe the failure time, a right censored or interval censored failure time as usually define.

The likelihood is then (Cox and Oakes, 1984)

$$
\begin{aligned}
L(\theta|\mathbf{t}, \mathbf{X}) = &\prod_{i \in \mathcal{U}} f(t_i|x_i, \theta) \prod_{i \in \mathcal{C}} \{1 - F(t_i^+|x_i, \theta\} \\
&\times \prod_{i \in \mathcal{I}} \{F(t_{iU}|x_i, \theta) - F(t_{iL}|x_i, \theta)\}
\end{aligned} \tag{1}
$$

4 Inference in the Mixtures of AFT Model Experts

We use MCMC with a data augmentation approach, where "true" survival times $\mathbf{w}$ and component membership $\mathbf{Z}$ (where $Z_{ij} = 1$ indicates observation i comes from component j) are considered latent variables.

Pseudo-observations w_i, $i = 1, \ldots, n$, are sampled. If t_i is not censored, $w_i = t_i$. If t_i is interval censored, w_i is sampled by solving

$$w_i = F^{-1}\left(F(t_{iL}|x_i, \theta) + u\left(F(t_{iU}|x_i, \theta) - F(t_{iL}|x_i, \theta)\right)\right) \tag{2}$$

where t_{iU} and t_{iL} are the upper and lower bounds for the time of occurrence of the event, and u is a uniform(0,1) random sample. If t_i is right censored, w_i is sampled by solving

$$w_i = F^{-1}\left(F(t_i^+|x_i, \theta) + u\left(1 - F(t_i^+|x_i, \theta)\right)\right) \tag{3}$$

where t_i^+ is the lower bound for the occurrence of t_i, and u is a uniform(0,1) random sample.

z_i, $i = 1 \ldots n$ are sampled, where z_i is a sample from a multinomial distribution with parameters $(1, h_{i1}, \ldots, h_{iJ})$, i.e.,

$$z_i \sim Multinomial(1, h_{i1}, \ldots, h_{iJ}) \tag{4}$$

with

$$h_{ij} = \frac{p_j(x_i, \rho_{1:J}) f_j(w_i|x_i, \beta_j)}{\sum_l^J p_l(x_i, \rho_{1:J}) f_l(w_i|x_i, \beta_l)}$$

then the augmented likelihood for the completely imputed data $\mathbf{w}$ and $\mathbf{z}$ is

$$L(\theta|\mathbf{z}, \mathbf{w}, \mathbf{t}, \mathbf{X}) = \prod_{i=1}^{n}\prod_{j=1}^{J}(p_j(\mathbf{x}_i, \rho) f_j(w_i|x_i, \beta_j, \sigma_j^2))^{z_{ij}}$$

We consider as prior distributions $(\beta_j|\sigma_j^2) \sim \mathcal{N}(b_{0j}, \sigma_j^2 B_{0j})$, and $\sigma_j^2 \sim \mathcal{IG}(n_{0j}/2, n_{0j}S_{0j}/2)$ independent for all $j = 1, \ldots, J$. And independently for $\rho \sim \mathcal{N}(\mathbf{0}, 1000 \times I)$. Then, the posterior distribution for the parameters is given by

$$\begin{aligned}\pi(\theta|\mathbf{z}, \mathbf{w}, \mathbf{t}, \mathbf{X}) \quad = \quad & \prod_{i=1}^{n}\prod_{j=1}^{J}(p_j(\mathbf{x}_i, \rho) f_j(w_i|x_i, \beta_j, \sigma_j^2))^{z_{ij}} \{\prod_{j=1}^{J} \phi(\beta_j|\mathbf{b}_{0j}, \sigma_j^2\mathbf{B}_{0j}) \times \mathcal{IG}(\sigma_j^2|n_{0j}/2, n_{0j}S_{0j}/2)\} \\ & \times \phi(\rho|\mathbf{0}, 1000 \times I)\end{aligned}$$

Algorithm:

1. Choose $\mathbf{w^0}$, $\mathbf{Z}^0$, $\theta^0 = \left\{\sigma_{1:J}^{2(0)}, \beta_{1:J}^0, \rho_{1:J}^0\right\}$
2. Sample $f(\mathbf{w}|\theta, \mathbf{Z})$ as equations (2) - (3)
3. Sample $f(\mathbf{Z}|\theta, \mathbf{w})$ as equation (4)
4. Sample $f(\theta|\mathbf{Z}, \mathbf{w})$, is explained next
5. Repeat 2-4 until convergence

Step 4, at this step we decide to blocking parameter θ into β, σ and δ, as follows

Note that, $\mathbf{z} = (z_1, \ldots, z_n)$ defines the group (or cluster, or component) from which the observation becomes.

Thus on each group (or cluster), the regression parameters β_j, $j = 1, \ldots, J$, are define by the sample $(\mathbf{Y}_j, \mathbf{X}_j)$, i.e. the observations that becomes from the j group (or cluster), then the full conditional posterior distribution for β_j is given by

$$\begin{aligned} \beta_j | \ldots &\sim \mathcal{N}(\mathbf{b}_{1j}, \sigma_j^2 \mathbf{B}_{1j}) \\ \sigma_j^2 | \ldots &\sim \mathcal{IG}(n_1/2, n_1 S_1/2) \end{aligned}$$

where $\mathcal{IG}$ denotes the inverse gamma distribution.

$$\begin{aligned} \mathbf{b}_{1j} &= \mathbf{B}_{1j}(\mathbf{B}_{0j}^{-1} b_{0j} + \mathbf{X}_j' \mathbf{Y}_j) \\ \mathbf{B}_{1j}^{-1} &= \mathbf{B}_{0j}^{-1} + \mathbf{X}_j' \mathbf{X}_j \\ n_1 &= n_{0j} + n_j \\ n_1 S_1 &= n_{0j} S_{0j} + (n_j - p) S_j^2 + (\hat{\beta}_j - \mathbf{b}_{0j})'(\mathbf{B}_{0j} + (\mathbf{X}_j' \mathbf{X}_j)^{-1})^{-1}(\hat{\beta}_j - \mathbf{b}_{0j}) \end{aligned}$$

and $\hat{\beta}_j = (\mathbf{X}_j' \mathbf{X}_j)^{-1} \mathbf{X}_j' \mathbf{Y}_j$, $S_j^2 = \mathbf{Y}_j'(\mathbf{I}_n - \mathbf{X}_j(\mathbf{X}_j' \mathbf{X}_j)^{-1} \mathbf{X}_j')/(n_j - p)$.

Note that, if $n_0 \to 0$ and $B_0^{-1} \to 0$ then the priori is non-informative.

For the ρ parameters a random walk Metropolis-Hastings step is used, due to the non-tractability of the full conditional posterior for this block of parameters.

The full conditional posterior distribution for ρ_j given the rest of parameters and observations is given by

$$\begin{aligned} \pi(\rho_j | \rho_{-j}, \ldots) &\propto \prod_i^n \prod_l^J (p(x_i, \rho_j) p(w_i | \ldots))^{z_{ij}} \\ &\propto \prod_i^n \frac{(\exp(x_i' \rho_j))^{z_{ij}}}{\sum_k \exp(x_i' \rho_k)} \\ &= \left(\prod_{i:z_{ij}=0} \frac{1}{\sum_k \exp(x_i' \rho_k)} \right) \left(\prod_{i:z_{ij}=1} \frac{\exp(x_i' \rho_j)}{\sum_k \exp(x_i' \rho_k)} \right) \end{aligned}$$