Web-based Supplementary Materials for “A Bayesian Semi-parametric Survival Model with Longitudinal Markers”

Song Zhang, Peter Müller and Kim-Anh Do

Web Appendix A. A Pseudo Prior Model

From $L_{vi0}$ with censored observations the two values taken by $\omega_{vi}$ lead to two models of different dimensions. If $\omega_{vi} = 0$, we have a model with $T_{vi}$ being a random parameter. In contrast, $T_{vi}$ is fixed at $T_{vi} = t_c$ if $\omega_{vi} = 1$. Such a change in dimension complicates posterior simulation (Green, 1995). We use the pseudo prior approach by Carlin and Chib (1995) to avoid this complication. In other words, we augment the smaller probability model under $\omega_{vi} = 1$ by defining a prior probability model for a hypothetical $T_{vi}$ (but keep $t_c$ in the regression for $y_{vi}$). The new variable $T_{vi}$ has no meaningful interpretation under $\omega_{vi} = 1$. It is only introduced to match the model dimensions. The augmented likelihood factor under the new model is

$$L^*_{vi0} = \{[y_{vi} | t_c, \Psi] p_v \pi_{vi}(T_{vi})\}^{\omega_{vi}} \cdot \{[y_{vi} | T_{vi}, \Psi](1 - p_v)g_v(T_{vi})I(T_{vi} > t_{vi})\}^{1 - \omega_{vi}}. \tag{1}$$

Here $\pi_{vi}(T_{vi})$ is a pseudo prior for $T_{vi}$ when $\omega_{vi} = 1$. It is a conveniently chosen linking density such that the two models implied by $\omega_{vi} = 0/1$ have the same dimension. The equivalence between $L^*_{vi0}$ and $L_{vi0}$ is obvious when $\omega_{vi} = 0$. When $\omega_{vi} = 1$, the equivalence can be verified by integrating (1) with regard to $T_{vi}$. As the name “pseudo prior” suggests, $\pi_{vi}(T_{vi})$ has no
effect on model inference. However, a poor choices of the linking density may lead to poor mixing of the posterior Markov chain Monte Carlo simulation. In our implementation we follow the recommendation of Carlin and Chib (1995) and base the specification of pseudo priors $\pi_{vi}(T_{vi})$ on a preliminary data analysis. First we fit a model without cure where $\omega_{vi} = 0$ for all subjects. Under this model there is no dimensional change. Then $\pi_{vi}(T_{vi})$ is specified to mimic the marginal posterior density of $T_{vi}$ under the simplified model. Specifically, we assume $\pi_{vi}(T_{vi})$ to be normal and match the first two moments. Finally, under the pseudo priors approach the marginal posterior distribution of $T_{vi}$ is meaningless. Only the posterior conditional density of $T_{vi}$ given $\omega_{vi} = 0$ is of interest.

Web Appendix B. The Polya Tree Prior

For reference, we give a brief review of PT models. More details can be found in Lavine (1992, 1994) and Mauldin et al. (1992). Let $\epsilon = \epsilon_1 \cdots \epsilon_m \in E^m$ denote a binary sequence of length $m$. For example, $E^1 = \{0, 1\}$ and $E^2 = \{00, 01, 10, 11\}$. The definition of the PT prior requires two parameters, a nested sequence of partitions $\Pi = \{B_0, B_1, B_{00}, B_{01}, \ldots, B_{\epsilon_0}, B_{\epsilon_1}, \ldots\}$ of the sample space $S$,

$$S = B_0 \cup B_1, \quad B_0 = B_{00} \cup B_{01}, \quad B_1 = B_{10} \cup B_{11}, \ldots, \quad B_\epsilon = B_{\epsilon_0} \cup B_{\epsilon_1}, \ldots,$$

and parameters $A = \{\alpha_0, \alpha_1, \alpha_{00}, \alpha_{01}, \ldots, \alpha_{\epsilon_0}, \alpha_{\epsilon_1}, \ldots\}$ that define a sequence of random variables $Y_{\epsilon_0} \sim Be(\alpha_{\epsilon_0}, \alpha_{\epsilon_1})$ and $Y_{\epsilon_1} = 1 - Y_{\epsilon_0}$, independently across $\epsilon$. We say that a random probability measure $G$ has a PT prior, $G \sim PT(\Pi, A)$, if the random probability $G(B_{\epsilon_0} \mid B_\epsilon)$ is defined by $G(B_{\epsilon_0} \mid B_\epsilon) \equiv Y_{\epsilon_0}$. This implies $G(B_{\epsilon_1, \ldots, \epsilon_m}) = \prod_{j=1}^{m} Y_{\epsilon_1, \ldots, \epsilon_j}$. We can center $G$ around a given distribution $\tilde{G}$, i.e., $E(G(B)) = \tilde{G}(B)$, by setting $\alpha_{\epsilon_0} = \alpha_{\epsilon_1}$ and taking the partition $\Pi$ at level $m$ to coincide with quantiles $\tilde{G}^{-1}(k/2^m)$, $k = 0, 1, \ldots, 2^m$. That is, for any $\epsilon \in E^m$,

$$B_\epsilon = (\tilde{G}^{-1}(k/2^m), \tilde{G}^{-1}((k + 1)/2^m))$$

(2)
for some $k$ in $\{0,1,\ldots,2^m-1\}$.

The family $\mathcal{A}$ determines how much $G$ varies around $\tilde{G}$. It has a similar role as the precision parameter in a Dirichlet process prior. Berger and Guglielmi (2001) considered a family of the form $\alpha_{\epsilon_1,\ldots,\epsilon_m} = c \cdot \rho(m)$, where $\rho(m) = m^2, m^3, 2^m, 4^m$ or $8^m$, and $c > 0$ is a constant. In general, any $\rho(m)$ such that $\sum_{m=1}^{\infty} \rho(m)^{-1} < \infty$ guarantees $G$ to be absolutely continuous. For example, $\rho(m) = m^{1+\eta}$ or $\rho(m) = (1 + \eta)^m$ for $\eta > 0$ satisfies the above condition.

A technically convenient property of PT priors is the conjugacy under random sampling. Let $n_\epsilon(T)$ be the number of elements of $T$ contained in $B_\epsilon$. The posterior distribution of $G$ given $T$ is again a PT, $G \mid T \sim PT(\Pi, \mathcal{A}')$, where $\mathcal{A}' = \{\alpha'_\epsilon\}$ with $\alpha'_\epsilon = \alpha_\epsilon + n_\epsilon(T)$.

Another useful property is the following closed form expression for the predictive density function of $(T_n \mid T_1, \ldots, T_{n-1})$, marginalized with respect to $G$. Let $T_{(-i)} = \{T_j : j \neq i\}$, let $\epsilon(j, T_i)$ denote the index $\epsilon_1 \cdots \epsilon_j \in E^j$ such that $T_i \in B_{\epsilon_1 \cdots \epsilon_j}$, and let $\tilde{g}(\cdot)$ be the density function of $\tilde{G}$. Assume that the partition $\Pi$ is specified as in (2), and $\mathcal{A}$ is specified such that for every $\epsilon \in E^m$, $\alpha_\epsilon = c \cdot m^2$. Define $M_i$ to be the smallest integer such that $n_\epsilon(M_i, T_i)(T_{(-i)}) = 0$. The marginal predictive distribution can be computed exactly:

$$[T_n \mid T_1, \ldots, T_{n-1}] = \left\{ \prod_{j=1}^{M_n} \frac{c j^2 + n_\epsilon(j, T_n)(T_{(-n)})}{2 c j^2 + n_\epsilon(j-1, T_n)(T_{(-n)})} \right\} 2^{M_i} \tilde{g}(T_n).$$

(3)

See, for example, Hanson and Johnson (2002).

The conditional cumulative probability marginalized with respect to $G$, $[T_n < t \mid T_1, \ldots, T_{n-1}]$ can also be evaluated exactly. We introduce the following notation. Let $M^*_i$ be the smallest integer such that $n_\epsilon(M^*_i, t)(T_{(-i)}) = 0$. Let $\mathcal{D}_i(t) = \{\epsilon : \epsilon \in E^{M^*_i} \text{ and } B_\epsilon < t\}$ be the set of indices for partitions defined at level $M^*_i$ and on the left of $t$. Here $B_\epsilon < t$ indicates that the upper bound of $B_\epsilon$ is smaller than $t$. For a partition $B_\epsilon$, $\epsilon \in \mathcal{D}_i(t)$, we define $\epsilon^*(j, B_\epsilon), j = 1, \ldots, M^*_i$, to be the sequence of indices such that $B_{\epsilon^*(1, B_\epsilon)} \supseteq B_{\epsilon^*(2, B_\epsilon)} \supseteq \cdots \supseteq B_{\epsilon^*(M^*_i, B_\epsilon)} = B_\epsilon$. We further define $B_i(t) = B_{\epsilon^*(M^*_i, B_\epsilon)} \cap (-\infty, t)$. Then the marginal cumulative
distribution function is

\[ T_n < t \mid T_1, \ldots, T_{n-1} = \sum_{\epsilon \in \mathcal{D}_n(t)} E(G(B_\epsilon)) + E(G(B_i(t))) \]

\[ = \sum_{\epsilon \in \mathcal{D}_n(t)} \prod_{j=1}^{M_\epsilon^*} \frac{c_j^2 + n_{\epsilon(j,B_\epsilon)}(T_{(-n)})}{2c_j^2 + n_{\epsilon(j-1,B_\epsilon)}(T_{(-n)})} \]

\[ + \prod_{j=1}^{M_\epsilon^*} \frac{c_j^2 + n_{\epsilon(j,t)}(T_{(-n)})}{2c_j^2 + n_{\epsilon(j-1,t)}(T_{(-n)})} \]

\[ M_{i}^* \tilde{G}(B_i(t)). \] (4)

The term \( \sum_{\epsilon \in \mathcal{D}_n(t)} E(G(B_\epsilon)) \) can be computed more efficiently when combining \( B_\epsilon, \epsilon \in \mathcal{D}_i(t) \), into partitions defined at higher levels of the Pólya tree. The marginalized conditional survival probability, \( [T_n > t \mid T_1, \ldots, T_{n-1}] \), can be computed in the same fashion. Expression (4) is useful in the computation of CPO.

### Web Appendix C. Posterior Sampling Scheme

Posterior MCMC simulation is built on sampling from the following conditional posterior distributions and other transition probabilities.

The simulation of the full conditional posterior distribution of \( \Psi \) depends on the prior model assumed. Different sampling strategies have been discussed by Gelman et al. (2003).

Under the pseudo priors setup, simulation of \( \omega_v^0 \) is straightforward. The full conditional posterior distribution of an unknown \( \omega_v \) is a Bernoulli(\( p_{vi}^* \)) with

\[ p_{vi}^* = \frac{[y_{vi} \mid t_{vi}, \Psi]p_v \pi_{vi}(T_{vi})}{[y_{vi} \mid t_{vi}, \Psi]p_v \pi_{vi}(T_{vi}) + [y_{vi} \mid T_{vi}, \Psi](1 - p_v)[T_{vi} \mid T_{v(-i)}^s]}. \]

Here \( T_{v(-i)}^s \) denotes the set of observed and unobserved TTP in the susceptible group except \( T_{vi} \). For censored subjects (\( d_{vi} = 0 \)), \( T_{vi} \) needs to be simulated. Given that \( \omega_v = 1 \), we simulate \( T_{vi} \) from the pseudo prior, i.e., \( T_{vi} \sim \pi_{vi}(T_{vi}) \). Given that \( \omega_v = 0 \), we simulate \( T_{vi} \) by Acceptance-Rejection sampling (Robert and Casella, 2003). The full conditional distribution of \( T_{vi} \) is proportional to \([y_{vi} \mid T_{vi}, \Psi]I(T_{vi} > t_{vi})[T_{vi} \mid T_{v(-i)}^s] \). Because \([y_{vi} \mid T_{vi}, \Psi] \) is bounded
and easy to evaluate, we can propose values based on \([T_v \mid T_v^{s(-i)}]I(T_v > t_v)\), and decide whether to accept or reject the proposal based on \([y_v \mid T_v, \Psi]\). The proposed values are generated from \([T_v \mid T_v^{s(-i)}]\) under the constraint that \(T_v > t_v\). The definition of \([T_v \mid T_v^{s(-i)}]\) is given in (3), which is the posterior predictive distribution under a PT prior. Lavine (1992) describes how to generate random samples from the posterior predictive distribution.

The full conditional distribution of \(p_v\) is proportional to

\[
p_v^{a_p + \sum_{i=1}^{n_v} \omega_{v_i} - 1} (1 - p_v)^{b_p + \sum_{i=1}^{n_v} (1 - \omega_{v_i}) - 1},
\]

which is again a Beta distribution.

**Web Appendix D. The Computation of CPO**

The computation of CPO differs for censored and uncensored subjects. We assume that \(d_{vi} = 0\) for \(i = 1, \ldots, n_v0\), and \(d_{vi} = 1\) for \(i = n_v0 + 1, \ldots, n_v\). First we derive CPO for uncensored cases, i.e., \(d_{vi} = 1\). Defining \(t_{v(-i)} = t_v/t_{vi}\) and integrating with respect to \(G_v\), we have

\[
CPO_{vi} = \int [y_v \mid t_v, \Psi][t_v \mid t_{v(-i)}]T_v^0(1 - p_v)[\Psi, T_v^0, p_v \mid Y_{v(-i)}, t_{(-v)}, d_{(-v)}]d\Psi dT_v^0 dp_v
\]

\[
= \int [y_v \mid t_v, \Psi][t_v \mid t_{v(-i)}]T_v^0(1 - p_v)[\Lambda \mid Y_{v(-i)}, t_{(-v)}, d_{(-v)}]d\Lambda.
\]  (5)

Note that when marginalized with regard to \(G_v\), the likelihood contribution from subject \((v, i)\) is \([y_v \mid t_v, \Psi][t_v \mid t_{v(-i)}, T_v^0](1 - p_v)\). We have the second equation because \([\Psi, T_v^0, p_v \mid Y_{v(-i)}, t_{(-v)}, d_{(-v)}]\) is the marginal distribution of \((\Psi, T_v^0, p_v)\) obtained from \([\Lambda \mid Y_{v(-i)}, t_{(-v)}, d_{(-v)}]\). Using the fact that

\[
[\Lambda \mid Y, t, d] \propto [y_v \mid t_v, \Psi][t_v \mid t_{v(-i)}]T_v^0(1 - p_v)[\Lambda \mid Y_{v(-i)}, t_{(-v)}, d_{(-v)}],
\]

we can evaluate (5) through an importance sampling scheme. The full posterior distribution, \([\Lambda \mid Y, t, d]\), serves as the importance sampling density, and the reciprocal of the likelihood
contribution from \((v, i)\) serves as the importance sampling weight. Specifically, we estimate \(CPO_{vi}\) by

\[
CPO_{vi} \approx \left\{ \frac{1}{K} \sum_{k=1}^{K} [y_{vi} | t_{vi}, \Psi] [t_{vi} | \Psi_{vi}] [t_{1v} T_{0v}^{0}(k)] (1 - p_{v}^{(k)}) \right\}^{-1},
\]

where \((\Psi^{(k)}, T_{0v}^{0(k)}, p_{v}^{(k)})\) is the \(k\)th sample from the full posterior distribution, \([\Lambda | Y, t, d]\), given all observations.

For censored cases, i.e., \(d_{vi} = 0\), the computation of CPO is more complicated. Integrating over \(G_{vi}\), the augmented likelihood factor is

\[
[y_{vi} | T_{vi}, \Psi] I(T_{vi} > t_{vi}) [T_{vi} | \omega_{vi}, t_{vi}, T_{0v}^{0}] [\omega_{vi} | p_{v}],
\]

where \(T_{0v}^{0(-i)} = T_{0v}/T_{vi}\) and \([T_{vi} | \omega_{vi}, t_{vi}, T_{0v}^{0(-i)}] = \{\delta(t_{c})\}^{\omega_{vi}} [T_{vi} | t_{vi}^{1}, T_{0v}^{0}]^{1-\omega_{vi}}\). Thus

\[
CPO_{vi} = \int [y_{vi} | T_{vi}, \Psi] I(T_{vi} > t_{vi}) [T_{vi} | \omega_{vi}, t_{vi}, T_{0v}^{0(-i)}] [\omega_{vi} | p_{v}]
\]

\[
\cdot [\Psi, T_{0v}^{0(-i)}, p_{v} | Y_{(-vi)}, t_{(-vi)}, d_{(-vi)}] d\Psi dT_{v}^{0} d\omega_{vi} dP_{v}.
\]

Note that \([\Psi, T_{0v}^{0(-i)}, p_{v} | Y_{(-vi)}, t_{(-vi)}, d_{(-vi)}]\) is obtained from \([\Lambda_{(-vi)} | Y_{(-vi)}, t_{(-vi)}, d_{(-vi)}]\) by marginalizing over parameters other than \((\Psi, T_{0v}^{0(-i)}, p_{v})\). Here \(\Lambda_{(-vi)}\) is the set of model parameters based on \((Y_{(-vi)}, t_{(-vi)}, d_{(-vi)})\), and \([\Lambda_{(-vi)} | Y_{(-vi)}, t_{(-vi)}, d_{(-vi)}]\) is the full posterior distribution as defined by Expression (2) in Zhang et al. (2008). Define

\[
P(T > t_{vi} | T_{0v}^{0(-i)}, p_{v}) = \int I(T_{vi} > t_{vi}) [T_{vi} | \omega_{vi}, t_{vi}^{1}, T_{0v}^{0(-i)}] [\omega_{vi} | p_{v}] d\omega_{vi} dT_{vi}
\]

\[
= \int I(T_{vi} > t_{vi}) \{p_{v}\delta(t_{c}) + (1 - p_{v})[T_{vi} | t_{vi}^{1}, T_{0v}^{0(-i)}]\} dT_{vi}
\]

\[
= p_{v} + (1 - p_{v})[T > t_{vi} | t_{vi}^{1}, T_{0v}^{0(-i)}],
\]

where \([T > t_{vi} | t_{vi}^{1}, T_{0v}^{0(-i)}]\) is the conditional survival probability of \(T_{vi}\) marginalized with respect to \(G_{vi}\), as is discussed in (4). We then define \(f(\Lambda)\) as the product of two density functions

\[
f(\Lambda) = \left\{ \frac{I(T_{vi} > t_{vi}) [T_{vi} | \omega_{vi}, t_{vi}^{1}, T_{0v}^{0(-i)}] [\omega_{vi} | p_{v}] \cdot [\Lambda_{(-vi)} | Y_{(-vi)}, t_{(-vi)}, d_{(-vi)}]}{P(T > t_{vi} | T_{0v}^{0(-i)}, p_{v})} \right\}.
\]
Thus we have

$$CPO_{vi} = \int [y_{vi} \mid T_{vi}, \Psi] P(T > t_{vi} \mid T_{v(-i)}^0, p_v) f(\Lambda) d\Lambda.$$ 

Since $[\Lambda \mid Y, t, d] \propto [y_{vi} \mid T_{vi}, \Psi] P(T > t_{vi} \mid T_{v(-i)}^0, p_v) f(\Lambda)$, an importance sampling scheme can be employed to evaluate CPO, with $[\Lambda \mid Y, t, d]$ being the importance sampling density and $\{(y_{vi} \mid T_{vi}, \Psi) P(T > t_{vi} \mid T_{v(-i)}^0, p_v)\}^{-1}$ being the importance sampling weight, i.e.,

$$CPO_{vi} \approx \left\{ \frac{1}{K} \sum_{k=1}^{K} \frac{1}{[y_{vi} \mid T_{vi}^{(k)}, \Psi^{(k)}] P(T > t_{vi} \mid T_{v(-i)}^{0(k)}, p_v^{(k)})} \right\}^{-1}. \quad (6)$$ 

Here $(\Psi^{(k)}, T_{vi}^{(k)}, T_{v(-i)}^{0(k)}, p_v^{(k)})$ are the kth sample from the full posterior distribution $[\Lambda \mid Y, t, d]$.

**Web Appendix E. Plots**

- Figure 1 plots the observed and fitted PSA profiles for four randomly selected patients.

- Figure 2 validates the survival and cure aspects of the proposed model based on subject specific martingale residuals (Barlow and Prentice, 1988; Therneau et al., 1990; Lin et al., 2002), which is defined by $e_{vi} = d_{vi} - H_{vi}$. Here $H_{vi}$ is the individual cumulative hazard up to $t_{vi}$. The residuals can be interpreted as the difference over $[0, t_{vi}]$ in the observed number of events and the expected number given the model. In general, the residuals are scattered horizontally over age (with three outliers), suggest that the proposed model is sufficient.

- Figure 3 shows the posterior variability of $G_v$ by plotting ten random samples from its posterior distribution.

- Figure 4 shows the initial drop and duration induced by the AA/CH treatments. A larger value of $\eta_v$ suggests a deeper initial drop in PSA level. On the other hand, the
larger the value of $\phi_{1v}$, the sooner the treatment effect wears out. We plot $l_v(t) = \eta_v[\exp(-\phi_{1v}t) - 1]$ in Figure 4.

References


Figure 1: The observed longitudinal profiles and fitted values of 4 randomly selected patients. The vertical axis indicates log($PSA + 1$), and the horizontal axis is age in years. Each point denotes a PSA measurement. The dotted lines plot fitted values of the longitudinal profiles. The vertical line marks the initiation of the AA/CH therapy.
Figure 2: The martingale residual for the survival model versus age. In general, the residuals are scattered horizontally over age.
Figure 3: Ten random samples from $[G_v | Y, t, d]$. The horizontal axis indicates years after treatment.
Figure 4: The initial slope and duration of treatment effect modeled by $\eta_v$ and $\phi_{1v}$. Here $t$ is the time in years from the start of treatment $v$ and $l_v(t) = \eta_v\{[\exp(-\phi_{1v}t) - 1]\}$.