A Bayesian Proportional-Hazards Model
In Survival Analysis

Stanley Sawyer — Washington University — August 24, 2004

1. Introduction. Suppose that a sample of \( n \) individuals has possible-censored survival times

\[
Y_1 \leq Y_2 \leq \ldots \leq Y_n \tag{1.1}
\]

Let \( \delta_i = 1 \) if the \( i^{th} \) time \( Y_i \) is an observed death and \( \delta_i = 0 \) if it was a right-censored event: That is, the individual was alive at time \( Y_i \), but was last seen at that time. If \( T_i \) (\( 1 \leq i \leq n \)) are the true survival or failure times, then \( Y_i = T_i \) if \( \delta_i = 1 \) and \( Y_i < T_i \) if \( \delta_i = 0 \), in which case the true failure time \( T_i \) is unknown.

We also assume \( d \)-dimensional covariate vectors \( X_1, X_2, \ldots, X_n \) for the \( n \) individuals in (1.1). The components of \( X_i \) might be age, income status, etc. The basic data for (1.1) is the set of triples \((Y_i, \delta_i, X_i)\) for \( 1 \leq i \leq n \). The most important statistical questions are connected with estimating the effect of the covariates \( X_i \) on the true survival times \( T_i \).

Let

\[
\tilde{Y}_1 < \tilde{Y}_2 < \ldots < \tilde{Y}_m \tag{1.2}
\]

be the distinct survival times in (1.1). At each time \( Y = \tilde{Y}_j \), let \( d_j \) be the number of observed deaths and \( a_j \) the number of censored events. Then \( n = \sum_{j=1}^{m} (d_j + a_j) \) is the total sample size and \( n_{\text{obs}} = \sum_{j=1}^{m} d_j = \sum_{i=1}^{n} \delta_i \) is the total number of observed deaths. The number of distinct observed death times is \( r = \sum_{j=1}^{m} I_{[d_j > 0]} \leq m \).

The basic statistical model that we describe below is essentially due to Kalbfleisch (1978). See Clayton (1991) and Ibrahim et al. (2001) for additional discussion and details, and Lee and Wang (2003) for an introduction to survival analysis. The model described below is nonparametric in flavor, but still allows tied survival-time data to be handled in a natural way. The likelihood formula that we derive below for tied data appears to be new. Previous work on this model has mostly assumed survival times (1.1) that are either without ties or else with grouped survival times (Kalbfleisch 1978, Ibrahim et al. 2001).

2. A Survival Model. Let \( Y \) be the true lifetime of a random individual with covariates \( X \). By definition, the survival function is

\[
S_X(t) = P_X(Y > t) = \exp(-H_X(t)) = \exp\left(-\int_0^t h_X(dy)\right) \tag{2.1}
\]
where \( H_X(t) \) is a right-continuous increasing function with \( H_X(0) = 0 \) and \( h_X(dy) \) is the related Lebesgue-Stieltjes measure. The function \( H_X(t) \) is one form of the cumulative hazard function and \( h_X(dy) \) the instantaneous hazard measure (or hazard rate). The Proportional Hazards assumption is

\[
h_X(dy) = e^\beta X h(dy) \quad \text{so that} \quad H_X(t) = e^\beta X H(t) \quad (2.2)
\]

for some \( d \)-dimensional vector of parameters \( \beta \), where \( \beta X \) in (2.2) is the dot product. One of the purposes of the model is to estimate \( \beta \) from the data and to test each component of \( \beta \) to find out whether that component of \( X \) has a statistically significant effect on the survival times \( Y \). A secondary goal is to estimate the baseline hazard density \( h(dy) \), which would allow us to estimate the expected survival time distribution \( S_X(t) \) for an individual whose covariates are \( X \), even if \( X \) is not among the covariate vectors \( X_i \) in the data.

In principle, the likelihood of the data \((Y_i, \delta_i, X_i)\) in (1.1) is

\[
L = \left( \prod_{\delta_i=0} P_{X_i} (Y > \tilde{Y}_i) \right) \left( \prod_{\delta_i=1} P_{X_i} (Y = \tilde{Y}_i) \right) \quad (2.3)
\]

where \( P_X(Y = \tilde{Y}_i) \) is with respect to some natural measure on the real line. Many inferential methods in statistics are based on finding the parameters that are the most likely for known data, in the sense of those parameters that have the largest value of \( L \).

To derive an explicit formula for (2.3), choose numbers \( \Delta_j > 0 \) such that \( \tilde{Y}_j + \Delta_j < \tilde{Y}_{j+1} - \Delta_{j+1} \) for all \( j \) and define the binned likelihood

\[
L_\Delta = \prod_{i=1}^n \left\{ \begin{array}{ll}
P_{X_i} (Y > \tilde{Y}_j + \Delta_j) & \text{if } \delta_i = 0 \\
P_{X_i} (\tilde{Y}_j - \Delta_j < Y \leq \tilde{Y}_j + \Delta_j) & \text{if } \delta_i = 1
\end{array} \right. \quad (2.4)
\]

By definition, the true lifetime \( T_i > \tilde{Y}_j \) for censored individuals with \( Y_i = \tilde{Y}_j \), so that (2.4) is the appropriate probability if the \( \Delta_j > 0 \) are sufficiently small. The likelihood (2.4) should be asymptotically proportional to (2.3) in the limit as \( \Delta_j \to 0 \).

We can write (2.4) in terms of the survival function \( S_X(t) \) in (2.1) as

\[
\prod_{i=1}^n \left\{ \begin{array}{ll}
S_{X_i} (\tilde{Y}_j + \Delta_j) & \text{if } \delta_i = 0 \\
S_{X_i} (\tilde{Y}_j - \Delta_j) - S_{X_i} (\tilde{Y}_j + \Delta_j) & \text{if } \delta_i = 1
\end{array} \right. = \prod_{i=1}^n \exp(- \int_0^{\tilde{Y}_j - \Delta_j} h_{X_i}(dy)) \left\{ \begin{array}{ll}
\exp\left(- \int_{\tilde{Y}_j - \Delta_j}^{\tilde{Y}_j + \Delta_j} h_{X_i}(dy) \right) & \text{if } \delta_i = 0 \\
1 - \exp\left(- \int_{\tilde{Y}_j - \Delta_j}^{\tilde{Y}_j + \Delta_j} h_{X_i}(dy) \right) & \text{if } \delta_i = 1
\end{array} \right.
\]
Define

\[ Z_j = \int_{Y_{j-1} + \Delta_{j-1}}^{Y_j - \Delta_j} h(dy) \quad \text{and} \quad Z_{j0} = \int_{Y_j - \Delta_j}^{Y_{j-1} + \Delta_{j-1}} h(dy) \quad (2.5) \]

Then

\[ \int_0^{Y_j - \Delta_j} h(dy) = Z_j + \sum_{k=1}^{j-1} (Z_k + Z_{k0}) = \sum_{k=1}^j Z_k + \sum_{k=1}^{j-1} Z_{k0} \]

so that

\[ \sum_{i=1}^n \int_0^{Y_j - \Delta_j} h_{X_i}(dy) = \sum_{i=1}^n e^{\beta X_i} \int_0^{Y_j - \Delta_j} h(dy) = \sum_{j=1}^{m} \left( \sum_{[Y_i = Y_j]} e^{\beta X_i} \right) \left( \sum_{k=1}^j Z_k + \sum_{k=1}^{j-1} Z_{k0} \right) \]

\[ = \sum_{j=1}^{m} Z_k \sum_{j=k}^{m} \left( \sum_{[Y_i = Y_j]} e^{\beta X_i} \right) + \sum_{k=1}^{m} Z_{k0} \sum_{j=k+1}^{m} \left( \sum_{[Y_i = Y_j]} e^{\beta X_i} \right) \]

\[ = \sum_{j=1}^{m} \left( Z_j R_j(\beta) + Z_{j0} R_{j+1}(\beta) \right) \quad (2.6) \]

In (2.6), \( R_j(\beta) \) is the risk sum

\[ R_j(\beta) = \sum_{k=j}^{m} \left( \sum_{[Y_i = Y_k]} e^{\beta X_i} \right) = \sum_{[Y_i \geq Y_j]} e^{\beta X_i} \quad (2.7) \]

corresponding to the individuals who are at risk immediately before time \( \tilde{Y}_j \).

We can then write the binned likelihood (2.4) as

\[ L_\Delta = \exp \left( - \sum_{j=1}^{m} \left( Z_j R_j(\beta) + Z_{j0} R_{j+1}(\beta) \right) \right) \prod_{j=1}^{m} \prod_{\left[ Y_i = \tilde{Y}_j \right]} e^{\beta X_i} \]

\[ \times \prod_{j=1}^{m} \prod_{\left[ Y_i = \tilde{Y}_j, \delta_i = 1 \right]} \left( 1 - \exp(-Z_{j0} e^{\beta X_i}) \right) \]

\[ = \exp \left( - \sum_{j=1}^{m} \left( Z_j R_j(\beta) + Z_{j0} R_{j}^0(\beta) + S_j(Z_{j0}, \beta) \right) \right) \quad (2.8) \]
where

\[ S_j(Z_{j0}, \beta) = \sum_{Y_i=\tilde{Y}_j, \delta_i=1} \log\left(1 - \exp\left(-Z_{j0} e^{\beta X_i}\right)\right) \tag{2.9} \]

is a sum over the observed deaths at times \( Y_i = \tilde{Y}_j \) and

\[ R_j^0(\beta) = \sum_{Y_i=\tilde{Y}_j, \delta_i=0} e^{\beta X_i} + \sum_{Y_i>\tilde{Y}_j} e^{\beta X_i} \tag{2.10} \]

is the risk sum for individuals who are at risk exactly immediately after time \( \tilde{Y}_j \).

3. A Gamma-process Prior for \( H(t) = \int_0^t h(dy) \). A useful way to estimate properties of the baseline hazard density \( h(dy) \) is to assume a parametric model for \( H(t) = \int_0^t h(dy) \) and then estimate the parameters involved. A useful parametric probability distribution for the set of increasing functions \( H(t) \) for \( t \geq 0 \) is the gamma process \( Z(t) \). This is a stochastic process with independent increments whose increments have the gamma distribution

\[ Z(t) - Z(s) \approx G\left(\theta(\alpha(t) - \alpha(s)), \lambda\right) \tag{3.1} \]

where \( \alpha(t) \) is some strictly-increasing function that is continuously differentiable for \( t > 0 \). In (3.1), \( Z \approx G(\theta, \lambda) \) means that \( Z \) is a random variable with the gamma probability density

\[ \frac{\lambda^\theta}{\Gamma(\theta)} x^{\theta-1} e^{-\lambda x} \quad \text{for} \quad 0 \leq z < \infty \]

Examples of \( \alpha(t) \) in (3.1) would be \( \alpha(t) = t \) or \( \alpha(t) = t^\sigma \) for some \( \sigma > 0 \). By (3.1),

\[ E(Z(t) - Z(s)) = \left(\theta(\alpha(t) - \alpha(s))\right)/\lambda = \mu(\alpha(t) - \alpha(s)) \quad \text{and} \]

\[ \text{Var}(Z(t) - Z(s)) = \left(\theta(\alpha(t) - \alpha(s))\right)/\lambda^2 = \mu(\alpha(t) - \alpha(s))/\lambda \tag{3.2} \]

for \( \mu = \theta/\lambda \).

If \( \alpha(t) = t \), \( E(Z(t)) = \mu t \) in (3.2), so that \( \alpha(t) = t \) corresponds to “noisy exponential” baseline survival times. Similarly, if \( \alpha(t) = t^\sigma \), then \( E(Z(t)) = \mu t^\sigma \), corresponding to “noisy Weibull” survival distributions. The function \( \alpha(t) \) is assumed fixed and \( \theta \) and \( \lambda \) are parameters to be estimated.
Given $\mu = \theta/\lambda$, $1/\lambda$ determines the variance of $H(t) = Z(t)$ about $E(H(t)) = \mu \alpha(t)$. Often $\theta$ or $\theta$ and $\lambda$ are given preassigned values to improve estimation.

The sample paths of the gamma process $Z(t)$ are, with probability one, strictly-increasing purely-discontinuous functions of $t$, although the probability that any preassigned value of $t$ is a jump is zero. This has the modeling advantage that tied survival-time values can occur with positive probability, even though the survival times themselves (not conditioned on the path $Z(t)$) have a continuous distribution, which means that any preassigned survival time has probability zero of being attained.

For any process $Z(t)$ with independent increments, the differences $Z_j, Z_{j0}$ in (2.5) are independent random variables. By (3.1), the $Z_j, Z_{j0}$ are independent random variables with gamma distributions

$$Z_j \approx \mathcal{G}(\theta W^\Delta_j, \lambda) \quad \text{where} \quad W^\Delta_j = \alpha(Y_j - \Delta_j) - \alpha(Y_{j-1} + \Delta_{j-1})$$

$$Z_{j0} \approx \mathcal{G}(\theta W^\Delta_{j0}, \lambda) \quad \text{where} \quad W^\Delta_{j0} = \alpha(Y_j + \Delta_j) - \alpha(Y_j - \Delta_j) \quad (3.3)$$

where we write $Y_j = \tilde{Y}_j$ for $\tilde{Y}_j$ in (1.2) for ease of notation.

If the $Z_j, Z_{j0}$ are considered parameters or “hidden variables” in the data $(Y_i, \delta_i, X_i)$ with the probability distribution (3.3), then the parameters $\theta, \lambda$ in (3.1) are considered hyperparameters. In a Bayesian framework, the hyperparameters themselves are given probability (or prior) distributions. In this case, we assume gamma prior distributions $\theta, \lambda \approx \mathcal{G}(\epsilon, \epsilon)$ for some small $\epsilon > 0$ ($\epsilon = 0.001$ is the most common choice) and an uninformative normal prior for each component $\beta_j$ of $\beta \in \mathbb{R}^d$, specifically that the prior distributions of $\beta_j$ are independent normal with means zero and standard deviation $1/\epsilon$ (Ibrahim et al. 2001). However, improper uniform priors for $\lambda$ and the $\beta_j$ would work just as well in this case.

4. The Full Likelihood $L$. Under these conditions, the full binned likelihood of the data, including the prior distributions for $Z_j, Z_{j0}$ and $\theta, \lambda, \beta$, corresponding to (2.8) is

$$L_\Delta = \frac{\epsilon^\epsilon}{\Gamma(\epsilon)} \theta^{\epsilon-1} e^{-\epsilon \theta} \frac{\epsilon^\epsilon}{\Gamma(\epsilon)} \lambda^{\epsilon-1} e^{-\epsilon \lambda} \prod_{a=1}^d \frac{1}{\sqrt{2\pi}} \exp\left(-\epsilon^2 \beta_a^2/2\right)$$

$$\times \prod_{j=1}^m \left( \frac{\lambda^{\theta W^\Delta_j}}{\Gamma(\theta W^\Delta_j)} Z_j^{\theta W^\Delta_j - 1} e^{-\lambda Z_j - Z_j R_j(\beta)} \right)$$

$$\times \frac{\lambda^{\theta W^\Delta_{j0}}}{\Gamma(\theta W^\Delta_{j0})} Z_{j0}^{\theta W^\Delta_{j0} - 1} e^{-\lambda Z_{j0} - Z_{j0} R_{j0}(\beta)} \prod_{[Y_i = \tilde{Y}_j, \delta_i = 1]} \left(1 - \exp(-Z_{j0} e^{\beta X_i})\right) \quad (4.1)$$
As each $\Delta_j \to 0$, 
\[ W_j^\Delta \to W_j = \alpha(Y_j) - \alpha(Y_{j-1}) > 0 \quad \text{and} \quad W_{j0}^\Delta \to 0 \quad (4.2) \]
for $Y_j = \bar{Y}_j$ as before. The expressions in the first line of (4.1) vary continuously as $W_j^\Delta \to W_j > 0$. As $W_{j0}^\Delta \to 0$, the $j^{th}$ factor in the second line in (4.1) is asymptotic to $C(Z_{j0})\Gamma(\theta W_{j0})^{-1} \sim C(Z_{j0})\theta \alpha'(\bar{Y}_j)\Delta_j$ for $Z_{j0} > 0$ and $C(Z_{j0}) > 0$. There are two cases for the asymptotic behavior of the $j^{th}$ factor in the second line in (4.1):

If $d_j = 0$, the $j^{th}$ factor has a delta-function singularity at $Z_{j0} = 0$ as $\Delta_j \to 0$ and $L_\Delta$ does not need to be rescaled. In this case, the factors in (4.1) with $Z_{j0}$ disappear in the limit as $\Delta_j \to 0$ (with $Z_{j0} = 0$).

If $d_j \geq 1$, the function $C(Z_{j0})$ is a bounded and continuous function of $Z_{j0}$ for $Z_{j0} \geq 0$ and the $j^{th}$ factor in (4.1) is asymptotic to $C(Z_{j0})\theta \alpha'(Y_j)\Delta_j$ as $\Delta_j \to 0$.

Thus, ignoring constants that depend on $\Delta_j$ for $d_j > 0$, the limit of $L_\Delta$ in (4.1) as $\max_j \Delta_j \to 0$ is the limiting full likelihood

\[
L = C \lambda^{e-1} e^{-\lambda} (\theta^{r-1} e^{-\theta}) \exp \left( -e^2 \sum_{a=1}^d 2 \right) 
\times \prod_{j=1}^m \left( \frac{\lambda W_j}{\Gamma(\theta W_j)} Z_j^{\theta W_j - 1} \exp \left( -Z_j (\lambda + R_j(\beta)) \right) \right) \quad (4.3) 
\times \prod_{[d_i \geq 1]} \exp \left( -Z_{j0} (\lambda + R_j(\beta)) \right) \left( \prod_{[Y_i = \bar{Y}_i, \delta_i = 1]} \left( 1 - \exp \left( -Z_{j0} e^\beta X_i \right) \right) \right)
\]

In (4.3), $C$ depends on $e$ and $\alpha'(Y_j)$ and $r$ is the number of distinct times $Y_j = \bar{Y}_j$ with $d_j \geq 1$. As mentioned earlier, inferences about which parameter values are relatively more likely are based on finding relatively larger values of $L$ in (4.3) for the data $(Y_i, \delta_i, X_i)$.

5. Estimating Parameters Using the Likelihood $L$. We estimate the parameters and hidden variables $(\theta, \lambda, Z_j, Z_{j0}, \beta)$ in (4.3) by using Markov Chain Monte Carlo methods (Metropolis et al. 1953, Hastings 1970, Gilks et al. 1996).

Specifically, we define a Markov Chain $Q_n$ that takes its values in the space of possible parameter vectors $(\theta, \lambda, Z_j, Z_{j0}, \beta)$ and which has a stationary or asymptotic distribution that is proportional to (4.3). This means that $Q_n$ spends most of its time where the likelihood (4.3) is the largest.
Mean or median values of components or functions of components of $Q_n$ can be used to provide estimates of the parameters affecting the true survival times $T_i$.

The Markov chain $Q_n$ proceeds by changing or updating each of the components of the vector $(\theta, \lambda, Z_j, Z_{j0}, \beta)$ in turn in a way that depends on the conditional probability distribution of that parameter value given the data and all the other parameters. We carry out these parameter changes or updates in the following way:

**Updating $\theta$:** Ignoring multiplicative constants and also ignoring factors in (4.3) that do not depend on $\theta$, the conditional density of $\theta$ given the data and the other parameters is

$$
\theta^{r+\epsilon-1}e^{-\epsilon\theta} \lambda^{\theta W} \prod_{j=1}^{m} \frac{Z_j^{\theta W_j}}{\Gamma(\theta W_j)} \quad \text{where} \quad W = \sum_{j=1}^{m} W_j
$$

(5.1)

for $W_j$ in (4.2). The density (5.1) is asymptotic to $C\theta^{r+m+\epsilon-1}$ as $\theta \to 0$ and decays faster than exponentially at infinity, and can be updated efficiently by one step of a Metropolis random walk (Metropolis et al. 1953).

Alternatively, the density (5.1) is a log-concave function of $\theta$, so that $\theta$ can be updated by a “Gibbs sampler” step that samples directly from the distribution (5.1) using one of the adaptive-rejection methods of Gilks and Wild (1992) or Gilks (1992). (See also Gilks et al. 1995.)

In general, a function $f(\theta)$ is called log-concave if $(d/d\theta)^2(\log f(\theta)) < 0$ for all $\theta$, or, more generally, if $(d/d\theta) \log f(\theta)$ is decreasing in $\theta$. The log-concavity of (5.1) follows from the identity

$$
\frac{d^2}{d\theta^2} \log \Gamma(\theta) = \text{Var}(\log \mathcal{G}(\theta, 1)) > 0
$$

(5.2)

where $\mathcal{G}(\theta, 1)$ represents a gamma-distributed random variable (as in (3.1)). *(Exercise: Prove (5.2).)*

**Updating $\lambda$:** Ignoring multiplicative constants and factors in (4.3) that do not depend on $\lambda$, the conditional density of $\lambda$ given the data and the other parameters is

$$
\lambda^{\theta W+\epsilon-1} \exp\left(-\lambda \left(\epsilon + \sum_{j=1}^{m} (Z_j + Z_{j0})\right)\right)
$$

(5.3)

where $Z_{j0} = 0$ if $d_j = 0$. This can be updated by a Gibbs sampler step by sampling from the gamma distribution

$$
\lambda \approx \mathcal{G}(\epsilon + \theta W, \epsilon + \sum_{j=1}^{m} (Z_j + Z_{j0})
$$
See Fishman (1995) for algorithms for generating gamma-distributed random variates. Two other good references for statistical computation and for scientific computing in general are Devroye (1986) and Press et al. (1992).

**Updating $Z_j$:** Ignoring multiplicative constants and factors that do not depend on $Z_j$, the conditional density of $Z_j$ given the other parameters is

$$Z_j^{\theta W_j - 1} e^{-Z_j(\lambda + R_j(\beta))}$$

Thus $Z_j$ can be updated by sampling from the gamma distribution

$$Z_j \approx G(\theta W_j, \lambda + R_j(\beta))$$

**Updating $Z_{j0}$ for $d_j \geq 1$:** Ignoring multiplicative constants and factors that do not depend on $Z_{j0}$, the conditional density of $Z_{j0}$ given the other parameters is

$$e^{-Z_{j0}(\lambda + R_{j0}(\beta))} \left( \prod_{Y_i = \tilde{Y}_j, \delta_i = 1} \left( 1 - \exp \left( -Z_{j0} e^{\beta X_i} \right) \right) \right)$$

The density (5.5) is normalizable in $Z_{j0}$ and can be updated by a one step of a Metropolis random walk. Unfortunately, the density (5.5) is not log-concave in $Z_{j0}$ due to the factor of $Z_{j0}$ in the denominator.

Alternatively, a more general sampling technique can be used for (5.5) that does not require log concavity (Gilks et al. 1995). This method, called “Metropolis-within-Gibbs” sampling, is equivalent to independence Metropolis-Hastings sampling (Gilks et al. 1996) using, as the proposal distribution, an approximation of the density (5.5) based on the method of Gilks (1992). If the density that is being approximated is log concave, the method reduces to the adaptive-rejection method of Gilks (1992).

Technically speaking, the term “Metropolis-within-Gibbs” is not quite correct, since independence sampling is not Metropolis sampling in the original sense. Metropolis et al. (1953) only described proposal distributions that are one step of a symmetric Markov chain. Independence sampling is contained in a generalization of Metropolis et al. (1953) due to Hastings (1970). The latter sampling scheme (or schemes) are usually called “Metropolis-Hastings” sampling.

Independence samplers can have extremely bad convergence properties if the proposal distribution is less singular or less heavy-tailed than the distribution being approximated (Gilks et al. 1996). In that case, a Metropolis random walk can be used instead.
Large values of $\beta X_i$ can cause numerical underflows and overflows in the exponential and risk sums in (5.5), but do not cause arbitrarily large values in the likelihood (5.5). Depending on the compiler, computer programs may have to be adjusted to avoid program crashes in evaluating $\exp(-Z_{j0} e^{\beta X_i})$ in (5.5) if $\beta X_i$ is large and positive. The term $\exp(-Z_{j0} e^{\beta X_i})$ in (5.5) can be set explicitly to be equal to zero in programming code if $Z_{j0} > 0$ and $\beta X_i > 500$ and equal to one if $\beta X_i < -500$. Most modern computers replace exponential underflows (that is, smaller positive values than the program can handle) by zero without a program warning or crash. If numerical underflows in exponentials can also cause program crashes, program adjustments may also have to be made if $Z_{j0} e^{\beta X_i}$ by itself is large.

**Updating $\beta$:** Ignoring multiplicative constants and factors that do not depend on $\beta$, the conditional density of $\beta$ in (4.3) given the data and other parameters is

$$
\exp\left(-\sum_{j=1}^{m} \left( Z_j R_j(\beta) + Z_{j0} R_0(\beta) - S_j(Z_{j0}, \beta) \right) - \frac{1}{2} \epsilon^2 \beta_a^2 \right)
$$

(5.6)

where

$$
S_j(Z_{j0}, \beta) = \sum_{Y_i = \tilde{Y}_j, \delta_i = 1} \log\left(1 - \exp(-Z_{j0} e^{\beta X_i})\right)
$$

is a sum over the observed deaths at times $Y_i = \tilde{Y}_j$ as in (2.9). If there are no observed deaths at time $Y_i = \tilde{Y}_j$, then $S_j(Z_{j0}, \beta) = 0$ and $Z_{j0} = 0$, and the second two terms in the sums in (5.6) do not appear.

Baring linear dependencies among the sample covariates, the conditional likelihood in (5.6) is normalizable in each component $\beta_j$, so that each $\beta_j$ can be updated efficiently by one step of a Metropolis random walk.

Alternatively, the density (5.6) is log-concave in $\beta_a$, so that Gibbs sampler updates can be made using the adaptive rejection methods of Gilks and Wild (1992) or Gilks (1992). Gilks has programming code in C on a Web site for carrying out Metropolis-within-Gibbs sampling that reduces to Gilks (1992) if a parameter is set. This C code can be used for non-Metropolis updates of $\theta$, $Z_{j0}$, and $\beta$.

6. **A Likelihood For $(\theta, \lambda, \beta)$**. The advantage of the Markov Chain Monte Carlo (MCMC) procedure of the previous section is that it also gives us information about the conditional distribution of the baseline cumulative hazards

$$
Z_j \approx H(Y_{j-}) - H(Y_{j-1}) \quad \text{and} \quad Z_{j0} \approx dH(Y_j) = h(dY_j)
$$
A Bayesian Proportional-Hazards Model

given the observed data. If we are primarily interested the parameters \((\theta, \lambda, \beta)\) and not in the baseline hazard density \(h(dy)\), the \(Z_j, Z_{j0}\) can be integrated out of the likelihood (4.3) to obtain a marginal likelihood that depends only on \((\theta, \lambda, \beta)\).

Evaluating the integrals \(\int L(Z_j) dZ_j\) in (4.3) in succession yields

\[
L = C \lambda^{r-1} e^{-\epsilon \lambda} \theta^r \prod_{j=1}^{m} \left( \frac{\lambda}{\lambda + R_j(\beta)} \right)^{\theta W_j} \times \exp \left( -\epsilon^2 \sum_{a=1}^{d} \beta_a^2 / 2 \right) \tag{6.1}
\]

\[
\times \prod_{[d_j \geq 1]} \exp \left( -Z_{j0} \left( \lambda + R_j^0(\beta) \right) \right) \left( \prod_{[Y_i = Y_j, \delta_i = 1]} \left( 1 - \exp \left( -Z_{j0} e^{\beta X_i} \right) \right) \right)
\]

While \(\lambda\) no longer has a simple gamma update, the parameter \(\theta\) now has a gamma update, specifically

\[
\theta \approx G \left( r + 1, \sum_{j=1}^{m} W_j \log \left( (\lambda + R_j(\beta))/\lambda \right) \right) \tag{6.2}
\]

The parameters \(Z_{j0}\) can be integrated by using the identity

\[
\int_{0}^{\infty} e^{-at} - e^{-bt} \frac{dt}{t} = \int_{0}^{\infty} \int_{a}^{b} e^{-\theta t} d\theta dt = \int_{a}^{b} \frac{d\theta}{\theta} = \log \frac{b}{a} \tag{6.3}
\]

for \(b > a > 0\). Thus if \(d_j = 1\), the \(j^{th}\) factor in the second line of (6.1) integrates to

\[
\log \left( \frac{\lambda + R_j^0(\beta) + e^{\beta X_i}}{\lambda + R_j(\beta)} \right) = \log \left( \frac{\lambda + R_j(\beta)}{\lambda + R_j^0(\beta)} \right)
\]

In particular, if \(d_j \leq 1\) for all \(j\), so that there are no ties among observed death times, then evaluating the integrals \(\int L(Z_{j0}) dZ_{j0}\) for \(d_j = 1\) in (6.1) leads to the more compact form

\[
L = L(\theta, \lambda, \beta) = C \lambda^{r-1} e^{-\epsilon \lambda} \theta^r \prod_{j=1}^{m} \left( \frac{\lambda}{\lambda + R_j(\beta)} \right)^{\theta W_j} \log \left( \frac{\lambda + R_j(\beta)}{\lambda + R_j^0(\beta)} \right) \tag{6.4}
\]

ignoring the prior terms in \(\beta\). If \(d_j = 0\), then \(R_j^0(\beta) = R_j(\beta)\) and the logarithmic factor does not appear. Analogous expressions can be found for \(d_j \geq 2\) by expanding the last product in (6.1) into a linear combination of differences of exponentials and applying (6.3).

The likelihood (6.4) no longer has information about the baseline hazards \(Z_j, Z_{j0}\), although the conditional density of \(Z_j, Z_{j0}\) is given by (5.4) and (5.5) if \(\beta\) is known precisely. See Kalbfleisch (1978) for a different derivation if \(d_j \leq 1\) for all \(j\).
7. The Posterior Distribution of the Hazard Function $H(t)$. (In Bayesian terminology, “posterior” means “conditional on the observed data for a given prior”.)

For any $j$ and any partition $(Y_{j-1}, Y_j) = \bigcup_{a=1}^{A_j} (Y_{j,a-1}, Y_{j,a})$ of $(Y_{j-1}, Y_j)$, define $Z_j = H(Y_j) - H(Y_{j-1}) = \sum_{a=1}^{A_j} Z_{ja}$ for $Z_{ja} = H(Y_{j,a}^-) - H(Y_{j,a-1})$.

The same argument as in (2.5) to (4.3) shows that the posterior distribution (4.3) is still valid with $Z_{ja}$ in place of $Z_j$, with of course $d_{ja} = 0$ unless there is an actual observed death at $Y_{j,a}$. This implies that the posterior distribution of the random variables $Z_{ja}$ is that they are independent gamma-distributed random variables with distributions

$$Z_{ja} \approx \mathcal{G} \left( \theta (\alpha(Y_{ja}) - \alpha(Y_{j,a-1})), \lambda + R_j(\beta) \right)$$

(7.1)

This in turn implies that, for each $j$, the posterior distribution of the process $Z(t) - Z(Y_{j-1}) = H(t) - H(Y_{j-1})$ for $Y_{j-1} < t < Y_j$ is that of a gamma process in $(Y_{j-1}, Y_j)$ with scale parameter $\lambda_j = \lambda + R_j(\beta)$, with jumps $Z(Y_{j+}) - Z(Y_{j-}) \approx Z_{j0}$ in the posterior distribution of $Z(t)$ at observed death times $Y_j$. As before, $Z_{j0} = 0$ if $d_j = 0$. If $d_j > 0$, $Z_{j0}$ has the density (5.5). (See also Kalbfleisch 1978 and Clayton 1991.)

8. Simulating Data for the Model in Sections 1–3. We can simulate survival data $(Y_i, \delta_i, X_i)$ for the model (2.1)–(2.2)–(3.1) as follows:

First, choose a sample size $n$, the number of covariates $d$, and, for $i \leq i \leq n$, covariates $X_i \in \mathbb{R}^d$. As in most regression models, these are assumed to be deterministic and are arbitrary. Choose arbitrary parameters values $\theta, \lambda > 0$ and risk parameters $\beta \in \mathbb{R}^d$. Also, choose a strictly-increasing continuously-differentiable function $\alpha(t)$ with $\alpha(0) = 0$, for example $\alpha(t) = t$.

The first goal is to define failure times $Y_i$ satisfying (2.1)–(2.2)–(3.1), that is

$$P(Y_i > t) = \exp(-H_{X_i}(t)) = \exp(-e^{\beta X_i} H(t)), \quad t \geq 0$$

(8.1)

where $H(t) = Z(t)$ is a realization of the gamma process

$$Z(t) \approx \mathcal{G}(\theta\alpha(t), \lambda) \approx (1/\lambda)\mathcal{G}(\theta\alpha(t), 1)$$

(8.2)

The final step will be to modify the construction so that some of the observations $Y_i$ can be censored.

The sample paths of $Z(t)$ are right-continuous with jumps in every time interval $(t_1, t_2)$ with $0 \leq t_1 < t_2$. This implies

$$P(Y_i > t) = P(Z(Y_i) > Z(t)) = \exp(-e^{\beta X_i} Z(t))$$
so that
\[ P(Z(Y_i) > s) = \exp \left( -e^{\beta X_i} s \right) \tag{8.3} \]
whenever \( s = Z(t) \). This suggests that \( Z(Y_i) \) might have an exponential distribution with mean \( e^{-\beta X_i} \), but this is not correct. In fact, given \( Z(t) \), the values of \( Z(Y_i) \) are restricted to the range of \( Z(t) \), which is the complement of an open dense set of real numbers since \( Z(t) \) is increasing with jumps in every open interval. This means that if the random variable \( Z(Y_i) \) has a probability distribution with a density \( g(s) \), then \( g(s) = 0 \) on an open dense set of real numbers \( s \). Thus \( Z(Y_i) \) cannot have a probability distribution with a continuous density.

If the variables \( Z(Y_i) \) were exponentially distributed, then we could simulate \( Y_i \approx Z^{-1}(Z_i) \) where \( Z_i \approx Z(Y_i) \) had a known distribution. However, we can do essentially the same even though the \( Z(Y_i) \) are not exponentially distributed.

Let \( Z_i \) be independent exponentially distribution random variables with mean \( e^{-\beta X_i} \), as incorrectly suggested for \( Z(Y_i) \) by (8.3). The \( Z_i \) can be simulated as
\[ Z_i \approx e^{-\beta X_i} \left( -\log(U_i) \right) \]
where \( U_i \) are independent uniforms for \( 0 \leq U_i \leq 1 \). Define
\[ Y_i = \min \{ t : Z(t) \geq Z_i \} \tag{8.4} \]
Then \( Y_i \leq t_2 \) if and only if \( Z(t_2) \geq Z_i \), so that
\[ P(Y_i > t) = P(Z(t) < Z_i) = \exp \left( -e^{\beta X_i} Z(t) \right) \tag{8.5} \]
which is exactly (8.1). If follows from (8.3) and (8.5) that \( P(Z(Y_i) \leq s) = P(Z_i \leq s) \) whenever \( s \) is a value attained by \( Z(t) \), but \( Z(Y_i) \) and \( Z_i \) have different probability distributions.

To simulate \( Y_i \) from (8.4), we need an approximate sample path of \( Z(t) \). Define independent gamma-distributed random variables
\[ Q_j \approx G(\theta \Delta(j, m), 1) \quad \text{for} \quad 1 \leq j \leq mT \tag{8.6} \]
where \( \Delta(j, m) = \alpha(j/m) - \alpha((j-1)/m) \) and \( m \) and \( T \) are large. In particular, \( \Delta(j, m) = 1/m \) if \( \alpha(t) = t \). In general, by (8.2) and (8.6),
\[ Z(k/m) \approx G(\theta \alpha(k/m), \lambda) \approx (1/\lambda)G(\theta \alpha(k/m), 1) \approx (1/\lambda) \sum_{j=1}^{k} Q_j \]
Thus we can simulate $Y_i$ in (8.4) by

$$Y_i = \min\{k/m : (1/\lambda) \sum_{j=1}^{k} Q_j \geq Z_i \} = \frac{1}{m} \min\{k : \sum_{j=1}^{k} Q_j \geq \lambda Z_i \}$$

or, equivalently, by

$$Y_i = \frac{1}{m} \min\{k : \sum_{j=1}^{k} Q_j \geq \tilde{Z}_i \} \quad \text{where} \quad (8.7)$$

$$\tilde{Z}_i \approx \lambda \exp(-\beta X_i)(-\log(U_i)) \approx \lambda Z_i$$

To include censoring, we define censoring times

$$Y_i^c = \frac{1}{m} \min\{k : \sum_{j=1}^{k} Q_j \geq \tilde{Z}_i^c \} \quad \text{for} \quad \tilde{Z}_i^c \approx \mu e^{-\beta X_i}(-\log(U_i))$$

in the same way for some constant $\mu > 0$. Define $\delta_i = 1$ (that is, the true failure time $Y_i = T_i$ is observed) if $Y_i < Y_i^c$ and $\delta_i = 0$ (that is, $Y_i < T_i$ and $Y_i$ is censored) if $Y_i^c < Y_i$. The last observed times (observed failure or censoring times) are

$$Y_i^o = \min\{Y_i, Y_i^c \} = \frac{1}{m} \min\{k : \sum_{j=1}^{k} Q_j \geq \tilde{Z}_i^o \}, \quad \tilde{Z}_i^o = \min\{\tilde{Z}_i, \tilde{Z}_i^c \}$$

(8.8)

In general, if $X_1$ and $X_2$ are independent exponentials with $E(X_1) = \mu_1$ and $E(X_2) = \mu_2$, then $X_3 = \min\{X_1, X_2\}$ is exponential with $E(X_3) = \mu_1\mu_2/(\mu_1 + \mu_2)$ and $P(X_1 < X_2) = \mu_2/(\mu_1 + \mu_2)$. Moreover, $X_3$ and the event $\{X_1 < X_2\}$ are independent. (Exercise: Prove these three statements.)

This implies that the triple $(Y_i^o, \delta_i, X_i)$ for $Y_i^o$ in (8.8) satisfies the conditions of the model (2.1)–(2.2)–(3.1) with $\lambda$ replaced by $\lambda\mu/(\lambda + \mu)$. Moreover, the variables $\delta_i = I(\tilde{Z}_i < \tilde{Z}_i^c)$ are independent with $P(\delta_i = 0) = P(\tilde{Z}_i < \tilde{Z}_i^c) = \lambda/(\lambda + \mu)$, and the $\delta_i$ are independent of $\tilde{Z}_i^o$.

This means that if we choose $\theta, \lambda$ and $0 < q < 1$ and define

$$Y_i = \frac{1}{m} \min\{k : \sum_{j=1}^{k} Q_j > \tilde{Z}_i \} \quad \text{for} \quad \tilde{Z}_i \approx \lambda(1-q)e^{-\beta X_i}(-\log(U_i))$$

and then, for each $i$, independently of the value of $Y_i$, call $Y_i$ censored ($\delta_i = 0$) with probability $q$ and observed ($\delta_i = 1$) with probability $1-q$, then $(Y_i, \delta_i, X_i)$ satisfy the conditions of Sections 1–3 with the original value of $\lambda$. (Exercise: Prove this. Note that if $\mu = ((1-q)/q)\lambda$, then $\lambda\mu/(\lambda + \mu) = (1-q)\lambda$ and $\lambda/(\lambda + \mu) = q$.)
A Bayesian Proportional-Hazards Model

References.


