Approximate Bayesian Inference for Cox Proportional Hazard Model

Ziang Zhang, Alex Stringer, Patrick Brown, Jamie Stafford

Department of Statistics, University of Toronto

2020-05-06

Outline

What is Survival Analysis?

Survival analysis refers to situations in which the response variable of interest is the time until the occurrence of a particular event.

For instances:

Time until death of patient with a specific disease,

Time to failure of a kind of light-bulbs,

What is Survival Analysis?

Survival analysis refers to situations in which the response variable of interest is the time until the occurrence of a particular event.

For instances:

Time until death of patient with a specific disease,
Time to failure of a kind of light-bulbs,

What is Survival Analysis?

Survival analysis refers to situations in which the response variable of interest is the time until the occurrence of a particular event.

For instances:

Time until death of patient with a specific disease,

Time to failure of a kind of light-bulbs,

What is Survival Analysis?

Survival analysis refers to situations in which the response variable of interest is the time until the occurrence of a particular event.

For instances:

Time until death of patient with a specific disease,
Time to failure of a kind of light-bulbs,

Example

We collect the information of 10 people (each with two kidneys) who have recovered from kidney infections such as gender, age, and type of infection on each kidney, and then record the recurrence times to infection of each kidney of each person. The study is one year long, and we hope to know whether a certain type of infection tend to recur faster conditional on the gender and age of patients.

1. Why don't we just do a linear regression?
   Reason: The response variable is not closed to normal. It is a positive random variable with possibly very skewed distribution (for example, exponential distribution).
2. Then why don't we just do a generalized linear regression instead?
   Reason: The exact realization of the response variable could be right-censored, so we don't know its exact value but only a lower bound for it. For example, the subject could be lost in the middle of the study, or the event never occur in the duration of the study.

1. Why don't we just do a linear regression?
   Reason: The response variable is not closed to normal. It is a positive random variable with possibly very skewed distribution (for example, exponential distribution).
2. Then why don't we just do a generalized linear regression instead?
   Reason: The exact realization of the response variable could be right-censored, so we don't know its exact value but only a lower bound for it. For example, the subject could be lost in the middle of the study, or the event never occur in the duration of the study.

1. Why don't we just do a linear regression?
   Reason: The response variable is not closed to normal. It is a positive random variable with possibly very skewed distribution (for example, exponential distribution).
2. Then why don't we just do a generalized linear regression instead?
   Reason: The exact realization of the response variable could be right-censored, so we don't know its exact value but only a lower bound for it. For example, the subject could be lost in the middle of the study, or the event never occur in the duration of the study.

1. Why don't we just do a linear regression?
   Reason: The response variable is not closed to normal. It is a positive random variable with possibly very skewed distribution (for example, exponential distribution).
2. Then why don't we just do a generalized linear regression instead?
   Reason: The exact realization of the response variable could be right-censored, so we don't know its exact value but only a lower bound for it. For example, the subject could be lost in the middle of the study, or the event never occur in the duration of the study.

Let Y_i denote the i-th response variable, with pdf $f(t)$ and cdf $F(t)$. Let y_i denote its realization or its censoring time.

Let δ_i is an indicator of whether the i-th observation is right-censored or not, so $\delta_i = 1$ means it is not right-censored so $Y_i = y_i$, and $\delta_i = 0$ means it is right-censored so $Y_i > y_i$.

In survival analysis, we have $\{(y_i, \delta_i) | i = 1, ..., n\}$ as our data for the response variable.

Let Y_i denote the i-th response variable, with pdf $f(t)$ and cdf $F(t)$. Let y_i denote its realization or its censoring time.

Let δ_i is an indicator of whether the i-th observation is right-censored or not, so $\delta_i = 1$ means it is not right-censored so $Y_i = y_i$, and $\delta_i = 0$ means it is right-censored so $Y_i > y_i$.

In survival analysis, we have $\{(y_i, \delta_i) | i = 1, ..., n\}$ as our data for the response variable.

Let Y_i denote the i-th response variable, with pdf $f(t)$ and cdf $F(t)$. Let y_i denote its realization or its censoring time.

Let δ_i is an indicator of whether the i-th observation is right-censored or not, so $\delta_i = 1$ means it is not right-censored so $Y_i = y_i$, and $\delta_i = 0$ means it is right-censored so $Y_i > y_i$.

In survival analysis, we have $\{(y_i, \delta_i) | i = 1, ..., n\}$ as our data for the response variable.

Two additional Definitions

Definition

Survival function of Y at time t is defined as:

$$S(t) = 1 - F(t) = \mathrm{P}(Y \geq t)$$

Definition (Finite Element Method)

Hazard function of Y at time t is defined as:

$$h(t) = \lim_{s \to 0} \frac{\mathrm{P}(t \leq Y \leq t+s | Y \geq t)}{s} = \frac{f(t)}{S(t)}$$

Two additional Definitions

Definition

Survival function of Y at time t is defined as:

$$S(t) = 1 - F(t) = \mathrm{P}(Y \geq t)$$

Definition (Finite Element Method)

Hazard function of Y at time t is defined as:

$$h(t) = \lim_{s\to 0} \frac{\mathrm{P}(t \leq Y \leq t+s | Y \geq t)}{s} = \frac{f(t)}{S(t)}$$

Definition

The Cox Proportional Hazard Model specifies the hazard function of the i-th observation Y_i as:

$$h_i(t) = h_0(t)\exp(\eta_i)$$

Here the η_i is the i-th additive linear predictor, for example: it could be $\beta_1 \text{age}_i + \beta_2 \text{sex}_i$ for a model with just two fixed effects.

The component $h_0(t)$ is a "universal" baseline hazard function that is the **same** for all the observations in our data set, but with an unspecified structure.

The main interest is for inference on components of η_i, because knowing those βs is enough for us to know the relative risk of one subject to another. The baseline hazard function $h_0(t)$ is often of secondary interest.

Definition

The Cox Proportional Hazard Model specifies the hazard function of the i-th observation Y_i as:

$$h_i(t) = h_0(t)\exp(\eta_i)$$

Here the η_i is the i-th additive linear predictor, for example: it could be $\beta_1 \text{age}_i + \beta_2 \text{sex}_i$ for a model with just two fixed effects.

The component $h_0(t)$ is a "universal" baseline hazard function that is the **same** for all the observations in our data set, but with an unspecified structure.

The main interest is for inference on components of η_i, because knowing those βs is enough for us to know the relative risk of one subject to another. The baseline hazard function $h_0(t)$ is often of secondary interest.

Definition

The Cox Proportional Hazard Model specifies the hazard function of the i-th observation Y_i as:

$$h_i(t) = h_0(t)\exp(\eta_i)$$

Here the η_i is the i-th additive linear predictor, for example: it could be $\beta_1 \text{age}_i + \beta_2 \text{sex}_i$ for a model with just two fixed effects.

The component $h_0(t)$ is a "universal" baseline hazard function that is the **same** for all the observations in our data set, but with an unspecified structure.

The main interest is for inference on components of η_i, because knowing those βs is enough for us to know the relative risk of one subject to another. The baseline hazard function $h_0(t)$ is often of secondary interest.

Definition

The Cox Proportional Hazard Model specifies the hazard function of the i-th observation Y_i as:

$$h_i(t) = h_0(t)\exp(\eta_i)$$

Here the η_i is the i-th additive linear predictor, for example: it could be $\beta_1 \text{age}_i + \beta_2 \text{sex}_i$ for a model with just two fixed effects.

The component $h_0(t)$ is a "universal" baseline hazard function that is the **same** for all the observations in our data set, but with an unspecified structure.

The main interest is for inference on components of η_i, because knowing those βs is enough for us to know the relative risk of one subject to another. The baseline hazard function $h_0(t)$ is often of secondary interest.

Note that the full-likelihood of CoxPH Model is:

$$
\begin{aligned}
L(y|\eta) &= \prod_{i=1}^{n} h_i(y_i)^{\delta_i} S_i(y_i) \\
&= \prod_{i=1}^{n} (h_0(y_i)\exp(\eta_i))^{\delta_i} S_i(y_i)
\end{aligned} \tag{1}
$$

To actually evaluate it, we need to know what function that $h_0(t)$ is, which unfortunately is rare in practice...

Note that the full-likelihood of CoxPH Model is:

$$
\begin{aligned}
L(y|\eta) &= \prod_{i=1}^{n} h_i(y_i)^{\delta_i} S_i(y_i) \\
&= \prod_{i=1}^{n} (h_0(y_i)\exp(\eta_i))^{\delta_i} S_i(y_i)
\end{aligned} \tag{1}
$$

To actually evaluate it, we need to know what function that $h_0(t)$ is, which unfortunately is rare in practice...

Partial likelihood of CoxPH Model can be viewed as a marginal likelihood for "observed ranks", and it doesn't require us to specify a baseline hazard function $h_0(t)$.

Denote the risk set of observation i as : $R_i := \{j \in 1 : n | y_j \geq y_i\}$, the partial likelihood can be written as:

$$\begin{aligned} L(y|\eta) &= \prod_{i=1}^{n} \left\{ \frac{h_i(y_i)}{\sum_{j \in R_i} h_j(y_i)} \right\}^{\delta_i} \\ &= \prod_{i=1}^{n} \left\{ \frac{\exp[\eta_i]}{\sum_{j \in R_i} \exp[\eta_j]} \right\}^{\delta_i} \end{aligned} \tag{2}$$

Note that the baseline hazard function $h_0(t)$ is cancelled out.

Partial likelihood of CoxPH Model can be viewed as a marginal likelihood for "observed ranks", and it doesn't require us to specify a baseline hazard function $h_0(t)$.
Denote the risk set of observation i as : $R_i := \{j \in 1 : n | y_j \geq y_i\}$, the partial likelihood can be written as:

$$
\begin{aligned}
L(y|\eta) &= \prod_{i=1}^{n} \left\{ \frac{h_i(y_i)}{\sum_{j \in R_i} h_j(y_i)} \right\}^{\delta_i} \\
&= \prod_{i=1}^{n} \left\{ \frac{\exp[\eta_i]}{\sum_{j \in R_i} \exp[\eta_j]} \right\}^{\delta_i}
\end{aligned} \tag{2}
$$

Note that the baseline hazard function $h_0(t)$ is cancelled out.

Partial likelihood of CoxPH Model can be viewed as a marginal likelihood for "observed ranks", and it doesn't require us to specify a baseline hazard function $h_0(t)$.
Denote the risk set of observation i as : $R_i := \{j \in 1 : n | y_j \geq y_i\}$, the partial likelihood can be written as:

$$\begin{aligned} L(y|\eta) &= \prod_{i=1}^{n} \left\{ \frac{h_i(y_i)}{\sum_{j \in R_i} h_j(y_i)} \right\}^{\delta_i} \\ &= \prod_{i=1}^{n} \left\{ \frac{\exp[\eta_i]}{\sum_{j \in R_i} \exp[\eta_j]} \right\}^{\delta_i} \end{aligned} \tag{2}$$

Note that the baseline hazard function $h_0(t)$ is cancelled out.

Partial likelihood of CoxPH Model can be viewed as a marginal likelihood for "observed ranks", and it doesn't require us to specify a baseline hazard function $h_0(t)$.
Denote the risk set of observation i as : $R_i := \left\{j \in 1 : n | y_j \geq y_i\right\}$, the partial likelihood can be written as:

$$
\begin{aligned}
L(y|\eta) &= \prod_{i=1}^{n} \left\{ \frac{h_i(y_i)}{\sum_{j \in R_i} h_j(y_i)} \right\}^{\delta_i} \\
&= \prod_{i=1}^{n} \left\{ \frac{\exp[\eta_i]}{\sum_{j \in R_i} \exp[\eta_j]} \right\}^{\delta_i}
\end{aligned}
\tag{2}
$$

Note that the baseline hazard function $h_0(t)$ is cancelled out.

Question: What if we are interested in Bayesian Inference of Cox Proportional Hazard Model?

- INLA with some extra conditions (Martino et al., 2011).
  - Can do approximate Bayesian inference on CoxPH model with the full likelihood.
  - It does not support the use of partial likelihood. To use the full likelihood, it approximates the baseline hazard function $h_0(t)$ with piece-wise constant functions, which may not work well when $h_0(t)$ is very complicated.
- Approximate Bayesian Inference for Case-crossover Model (Stringer et al., 2020).
  - It does support the use of partial likelihood in its inference.
  - It can only do the inference for case-crossover model. The partial likelihood of case-crossover model has similar form with Cox PH model, but is simpler than Cox PH model in general.

Question: What if we are interested in Bayesian Inference of Cox Proportional Hazard Model?

- INLA with some extra conditions (Martino et al., 2011).
  - Can do approximate Bayesian inference on CoxPH model with the full likelihood.
  - It does not support the use of partial likelihood. To use the full likelihood, it approximates the baseline hazard function $h_0(t)$ with piece-wise constant functions, which may not work well when $h_0(t)$ is very complicated.
- Approximate Bayesian Inference for Case-crossover Model (Stringer et al., 2020).
  - It does support the use of partial likelihood in its inference.
  - It can only do the inference for case-crossover model. The partial likelihood of case-crossover model has similar form with Cox PH model, but is simpler than Cox PH model in general.

Question: What if we are interested in Bayesian Inference of Cox Proportional Hazard Model?

- INLA with some extra conditions (Martino et al., 2011).
  - Can do approximate Bayesian inference on CoxPH model with the full likelihood.
  - It does not support the use of partial likelihood. To use the full likelihood, it approximates the baseline hazard function $h_0(t)$ with piece-wise constant functions, which may not work well when $h_0(t)$ is very complicated.
- Approximate Bayesian Inference for Case-crossover Model (Stringer et al., 2020).
  - It does support the use of partial likelihood in its inference.
  - It can only do the inference for case-crossover model. The partial likelihood of case-crossover model has similar form with Cox PH model, but is simpler than Cox PH model in general.

Question: What if we are interested in Bayesian Inference of Cox Proportional Hazard Model?

- INLA with some extra conditions (Martino et al., 2011).
  - Can do approximate Bayesian inference on CoxPH model with the full likelihood.
  - It does not support the use of partial likelihood. To use the full likelihood, it approximates the baseline hazard function $h_0(t)$ with piece-wise constant functions, which may not work well when $h_0(t)$ is very complicated.
- Approximate Bayesian Inference for Case-crossover Model (Stringer et al., 2020).
  - It does support the use of partial likelihood in its inference.
  - It can only do the inference for case-crossover model. The partial likelihood of case-crossover model has similar form with Cox PH model, but is simpler than Cox PH model in general.

Question: What if we are interested in Bayesian Inference of Cox Proportional Hazard Model?

- INLA with some extra conditions (Martino et al., 2011).
  - Can do approximate Bayesian inference on CoxPH model with the full likelihood.
  - It does not support the use of partial likelihood. To use the full likelihood, it approximates the baseline hazard function $h_0(t)$ with piece-wise constant functions, which may not work well when $h_0(t)$ is very complicated.
- Approximate Bayesian Inference for Case-crossover Model (Stringer et al., 2020).
  - It does support the use of partial likelihood in its inference.
  - It can only do the inference for case-crossover model. The partial likelihood of case-crossover model has similar form with Cox PH model, but is simpler than Cox PH model in general.

Question: What if we are interested in Bayesian Inference of Cox Proportional Hazard Model?

- INLA with some extra conditions (Martino et al., 2011).
  - Can do approximate Bayesian inference on CoxPH model with the full likelihood.
  - It does not support the use of partial likelihood. To use the full likelihood, it approximates the baseline hazard function $h_0(t)$ with piece-wise constant functions, which may not work well when $h_0(t)$ is very complicated.
- Approximate Bayesian Inference for Case-crossover Model (Stringer et al., 2020).
  - It does support the use of partial likelihood in its inference.
  - It can only do the inference for case-crossover model. The partial likelihood of case-crossover model has similar form with Cox PH model, but is simpler than Cox PH model in general.

Question: What if we are interested in Bayesian Inference of Cox Proportional Hazard Model?

- INLA with some extra conditions (Martino et al., 2011).
  - Can do approximate Bayesian inference on CoxPH model with the full likelihood.
  - It does not support the use of partial likelihood. To use the full likelihood, it approximates the baseline hazard function $h_0(t)$ with piece-wise constant functions, which may not work well when $h_0(t)$ is very complicated.
- Approximate Bayesian Inference for Case-crossover Model (Stringer et al., 2020).
  - It does support the use of partial likelihood in its inference.
  - It can only do the inference for case-crossover model. The partial likelihood of case-crossover model has similar form with Cox PH model, but is simpler than Cox PH model in general.

Let's have a brief revisit of the general Cox Proportional Hazard Model (with some small modifications):

$$h_i(t) = h_0(t)\exp(\eta_i)$$

where the additive linear predictor η_i can be generally defined as:

$$\eta_i = x_i^T \beta + \gamma(u_i) + \epsilon_i$$

- x is a vector of fixed effect covariates.
- u is the smoothing covariate, and γ is its corresponding smoothing function that we want to make inference on.
- $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ is an auxiliary variable to make the computation easier, and τ^{-1} is very small.

Unlike the case-crossover model, it is also possible to include a random intercepts for each subject, which is called "frailty" between subjects in survival analysis because of the form of its partial likelihood (Vaupel et al., 1979).

Let's have a brief revisit of the general Cox Proportional Hazard Model (with some small modifications):

$$h_i(t) = h_0(t)\exp(\eta_i)$$

where the additive linear predictor η_i can be generally defined as:

$$\eta_i = x_i^T\beta + \gamma(u_i) + \epsilon_i$$

- x is a vector of fixed effect covariates.
- u is the smoothing covariate, and γ is its corresponding smoothing function that we want to make inference on.
- $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ is an auxiliary variable to make the computation easier, and τ^{-1} is very small.

Unlike the case-crossover model, it is also possible to include a random intercepts for each subject, which is called "frailty" between subjects in survival analysis because of the form of its partial likelihood (Vaupel et al., 1979).

Let's have a brief revisit of the general Cox Proportional Hazard Model (with some small modifications):

$$h_i(t) = h_0(t)\exp(\eta_i)$$

where the additive linear predictor η_i can be generally defined as:

$$\eta_i = x_i^T\beta + \gamma(u_i) + \epsilon_i$$

- x is a vector of fixed effect covariates.
- u is the smoothing covariate, and γ is its corresponding smoothing function that we want to make inference on.
- $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ is an auxiliary variable to make the computation easier, and τ^{-1} is very small.

Unlike the case-crossover model, it is also possible to include a random intercepts for each subject, which is called "frailty" between subjects in survival analysis because of the form of its partial likelihood (Vaupel et al., 1979).

Let's have a brief revisit of the general Cox Proportional Hazard Model (with some small modifications):

$$h_i(t) = h_0(t)\exp(\eta_i)$$

where the additive linear predictor η_i can be generally defined as:

$$\eta_i = x_i^T\beta + \gamma(u_i) + \epsilon_i$$

- x is a vector of fixed effect covariates.
- u is the smoothing covariate, and γ is its corresponding smoothing function that we want to make inference on.
- $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ is an auxiliary variable to make the computation easier, and τ^{-1} is very small.

Unlike the case-crossover model, it is also possible to include a random intercepts for each subject, which is called "frailty" between subjects in survival analysis because of the form of its partial likelihood (Vaupel et al., 1979).

Let's have a brief revisit of the general Cox Proportional Hazard Model (with some small modifications):

$$h_i(t) = h_0(t)\exp(\eta_i)$$

where the additive linear predictor η_i can be generally defined as:

$$\eta_i = x_i^T\beta + \gamma(u_i) + \epsilon_i$$

- x is a vector of fixed effect covariates.
- u is the smoothing covariate, and γ is its corresponding smoothing function that we want to make inference on.
- $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \tau^{-1})$ is an auxiliary variable to make the computation easier, and τ^{-1} is very small.

Unlike the case-crossover model, it is also possible to include a random intercepts for each subject, which is called "frailty" between subjects in survival analysis because of the form of its partial likelihood (Vaupel et al., 1979).

For the inference of smoothing function, we implemented second-order random walk model (Lindgren and Rue, 2008)

- The smoothing covariate u is first discretized into m disjoint bins : $\{[a_1, a_2], [a_2, a_3]...., [a_m, a_{m+1}]\}$. Denote the meddle points of these bins as $\{u_1, ..., u_m\}$.
- Assume the smoothing function γ is piece-wise constant within each bin, so the inference on γ becomes inference on $\Gamma = \{\gamma(u_1), \gamma(u_2), ..., \gamma(u_m)\}$.
- Put a second-order random walk prior on Γ, so $(\gamma(u_{i+1}) - \gamma(u_i)) - (\gamma(u_i) - \gamma(u_{i-1})) = \gamma(u_{i-1}) - 2\gamma(u_i) + \gamma(u_{i+1}) \overset{iid}{\sim} \mathrm{N}(0, \sigma_u)$

For the inference of smoothing function, we implemented second-order random walk model (Lindgren and Rue, 2008)

- The smoothing covariate u is first discretized into m disjoint bins : $\{[a_1, a_2], [a_2, a_3]...., [a_m, a_{m+1}]\}$. Denote the meddle points of these bins as $\{u_1, ..., u_m\}$.
- Assume the smoothing function γ is piece-wise constant within each bin, so the inference on γ becomes inference on $\Gamma = \{\gamma(u_1), \gamma(u_2), ..., \gamma(u_m)\}$.
- Put a second-order random walk prior on Γ, so $(\gamma(u_{i+1}) - \gamma(u_i)) - (\gamma(u_i) - \gamma(u_{i-1})) = \gamma(u_{i-1}) - 2\gamma(u_i) + \gamma(u_{i+1}) \overset{iid}{\sim} \mathrm{N}(0, \sigma_u)$

For the inference of smoothing function, we implemented second-order random walk model (Lindgren and Rue, 2008)

- The smoothing covariate u is first discretized into m disjoint bins : $\{[a_1, a_2], [a_2, a_3]...., [a_m, a_{m+1}]\}$. Denote the meddle points of these bins as $\{u_1, ..., u_m\}$.
- Assume the smoothing function γ is piece-wise constant within each bin, so the inference on γ becomes inference on $\Gamma = \{\gamma(u_1), \gamma(u_2), ..., \gamma(u_m)\}$.
- Put a second-order random walk prior on Γ, so $(\gamma(u_{i+1}) - \gamma(u_i)) - (\gamma(u_i) - \gamma(u_{i-1})) = \gamma(u_{i-1}) - 2\gamma(u_i) + \gamma(u_{i+1}) \overset{iid}{\sim} \mathrm{N}(0, \sigma_u)$

For the inference of smoothing function, we implemented second-order random walk model (Lindgren and Rue, 2008)

- The smoothing covariate u is first discretized into m disjoint bins : $\{[a_1, a_2], [a_2, a_3]...., [a_m, a_{m+1}]\}$. Denote the meddle points of these bins as $\{u_1, ..., u_m\}$.
- Assume the smoothing function γ is piece-wise constant within each bin, so the inference on γ becomes inference on $\Gamma = \{\gamma(u_1), \gamma(u_2), ..., \gamma(u_m)\}$.
- Put a second-order random walk prior on Γ, so $(\gamma(u_{i+1}) - \gamma(u_i)) - (\gamma(u_i) - \gamma(u_{i-1})) = \gamma(u_{i-1}) - 2\gamma(u_i) + \gamma(u_{i+1}) \overset{iid}{\sim} \mathrm{N}(0, \sigma_u)$

Then, also put a joint Gaussian prior on β such that:
$\beta \sim \mathrm{N}(0, \Sigma_\beta)$.
Let $W = (\beta, \Gamma, \eta)$, then by construction, W follows a joint Gaussian as well.
The hyper-parameter θ is defined as $-2\log(\sigma_u)$, which controls the level of smoothness. The prior put on this is not necessary Gaussian.

For the approximation method, we adopt the inferential methodology of Stringer et al. (2020).
The objects of inferential interest are the posteriors:

$$\begin{aligned}\pi(W_j|\boldsymbol{Y}) &= \int\int \pi(\boldsymbol{W}|\boldsymbol{Y},\theta)\pi(\theta|\boldsymbol{Y})d\boldsymbol{W}_{-j}d\theta,\\ \pi(\theta|\boldsymbol{Y}) &= \frac{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}}{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}d\theta}\end{aligned} \tag{3}$$

1. Approximate $\pi(\theta^k|\boldsymbol{Y}) \approx \tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})$, a **Laplace approximation** (Tierney and Kadane, 1986),
2. Approximate $\pi(\boldsymbol{W}|\boldsymbol{Y},\theta^k) \approx \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)$, a **Gaussian approximation**. This also gives a univariate Gaussian approximation $\tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k) = \int \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)d\boldsymbol{W}_{-j}$.

Choose a **grid** and corresponding weights $\{\theta^k, \Delta^k : k \in [K]\}$. Then approximate $\pi(W_j|\boldsymbol{Y}) \approx \sum_{k=1}^{K} \tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k)\tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})\Delta^k$.

For the approximation method, we adopt the inferential methodology of Stringer et al. (2020).
The objects of inferential interest are the posteriors:

$$\begin{aligned} \pi(W_j|\boldsymbol{Y}) &= \int\int \pi(\boldsymbol{W}|\boldsymbol{Y},\theta)\pi(\theta|\boldsymbol{Y})d\boldsymbol{W}_{-j}d\theta, \\ \pi(\theta|\boldsymbol{Y}) &= \frac{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}}{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}d\theta} \end{aligned} \tag{3}$$

1. Approximate $\pi(\theta^k|\boldsymbol{Y}) \approx \tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})$, a **Laplace approximation** (Tierney and Kadane, 1986),
2. Approximate $\pi(\boldsymbol{W}|\boldsymbol{Y},\theta^k) \approx \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)$, a **Gaussian approximation**. This also gives a univariate Gaussian approximation $\tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k) = \int \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)d\boldsymbol{W}_{-j}$.

Choose a **grid** and corresponding weights $\{\theta^k, \Delta^k : k \in [K]\}$. Then approximate $\pi(W_j|\boldsymbol{Y}) \approx \sum_{k=1}^{K} \tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k)\tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})\Delta^k$.

For the approximation method, we adopt the inferential methodology of Stringer et al. (2020).
The objects of inferential interest are the posteriors:

$$\begin{aligned} \pi(W_j|\boldsymbol{Y}) &= \int\int \pi(\boldsymbol{W}|\boldsymbol{Y},\theta)\pi(\theta|\boldsymbol{Y})d\boldsymbol{W}_{-j}d\theta, \\ \pi(\theta|\boldsymbol{Y}) &= \frac{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}}{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}d\theta} \end{aligned} \tag{3}$$

1. Approximate $\pi(\theta^k|\boldsymbol{Y}) \approx \tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})$, a **Laplace approximation** (Tierney and Kadane, 1986),
2. Approximate $\pi(\boldsymbol{W}|\boldsymbol{Y},\theta^k) \approx \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)$, a **Gaussian approximation**. This also gives a univariate Gaussian approximation $\tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k) = \int \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)d\boldsymbol{W}_{-j}$.

Choose a **grid** and corresponding weights $\{\theta^k, \Delta^k : k \in [K]\}$. Then approximate $\pi(W_j|\boldsymbol{Y}) \approx \sum_{k=1}^{K} \tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k)\tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})\Delta^k$.

For the approximation method, we adopt the inferential methodology of Stringer et al. (2020).
The objects of inferential interest are the posteriors:

$$\begin{aligned} \pi(W_j|\boldsymbol{Y}) &= \int\int \pi(\boldsymbol{W}|\boldsymbol{Y},\theta)\pi(\theta|\boldsymbol{Y})d\boldsymbol{W}_{-j}d\theta, \\ \pi(\theta|\boldsymbol{Y}) &= \frac{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}}{\int \pi(\boldsymbol{W},\theta,\boldsymbol{Y})d\boldsymbol{W}d\theta} \end{aligned} \tag{3}$$

1. Approximate $\pi(\theta^k|\boldsymbol{Y}) \approx \tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})$, a **Laplace approximation** (Tierney and Kadane, 1986),
2. Approximate $\pi(\boldsymbol{W}|\boldsymbol{Y},\theta^k) \approx \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)$, a **Gaussian approximation**. This also gives a univariate Gaussian approximation $\tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k) = \int \tilde{\pi}_G(\boldsymbol{W}|\boldsymbol{Y},\theta^k)d\boldsymbol{W}_{-j}$.

Choose a **grid** and corresponding weights $\left\{\theta^k, \Delta^k : k \in [K]\right\}$. Then approximate $\pi(W_j|\boldsymbol{Y}) \approx \sum_{k=1}^{K} \tilde{\pi}_G(W_j|\boldsymbol{Y},\theta^k)\tilde{\pi}_{LA}(\theta^k|\boldsymbol{Y})\Delta^k$.

The approximation methodology is **fast** for data set with small to medium size.

The **Gaussian approximation** requires repeated high-dimensional optimizations. The objective function is **convex** but has a **dense Hessian**, so we use **trust region** optimization with **quasi-Newton method** (Braun, 2014) to do this **fast** and **stable**.

The quasi Newton method (SR1: Symmetric Rank One) **avoids** the evaluation of the actual Hessian matrix **in each iteration of the optimization**, and therefore reduces the computational complexity during the optimization. But we still **need** the value of the Hessian matrix **at each optimal points** for the approximation.

The computations which need to be done for multiple θ^k are done in **parallel**.

The approximation methodology is **fast** for data set with small to medium size.

The **Gaussian approximation** requires repeated high-dimensional optimizations. The objective function is **convex** but has a **dense Hessian**, so we use **trust region** optimization with **quasi-Newton method** (Braun, 2014) to do this **fast** and **stable**.

The quasi Newton method (SR1: Symmetric Rank One) **avoids** the evaluation of the actual Hessian matrix **in each iteration of the optimization**, and therefore reduces the computational complexity during the optimization. But we still **need** the value of the Hessian matrix **at each optimal points** for the approximation.

The computations which need to be done for multiple θ^k are done in **parallel**.

The approximation methodology is **fast** for data set with small to medium size.

The **Gaussian approximation** requires repeated high-dimensional optimizations. The objective function is **convex** but has a **dense Hessian**, so we use **trust region** optimization with **quasi-Newton method** (Braun, 2014) to do this **fast** and **stable**.

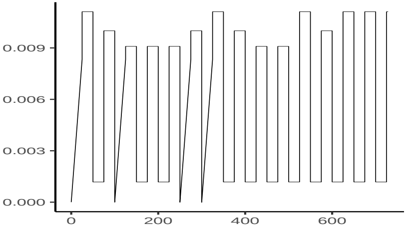
The quasi Newton method (SR1: Symmetric Rank One) **avoids** the evaluation of the actual Hessian matrix **in each iteration of the optimization**, and therefore reduces the computational complexity during the optimization. But we still **need** the value of the Hessian matrix **at each optimal points** for the approximation.

The computations which need to be done for multiple θ^k are done in **parallel**.

The approximation methodology is **fast** for data set with small to medium size.
The **Gaussian approximation** requires repeated high-dimensional optimizations. The objective function is **convex** but has a **dense Hessian**, so we use **trust region** optimization with **quasi-Newton method** (Braun, 2014) to do this **fast** and **stable**.
The quasi Newton method (SR1: Symmetric Rank One) **avoids** the evaluation of the actual Hessian matrix **in each iteration of the optimization**, and therefore reduces the computational complexity during the optimization. But we still **need** the value of the Hessian matrix **at each optimal points** for the approximation.
The computations which need to be done for multiple θ^k are done in **parallel**.

To show the accuracy of our approach compared to INLA when the baseline hazard function is non-smooth, we did the following simulation:

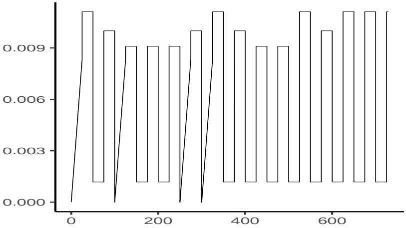
- We simulate $\mathbf{N} = 400$ independent observations from a CoxPH Model with the following baseline hazard function:



- For simplicity, we assume the linear predictor is $\eta_i = \gamma(u_i) + \epsilon_i$, with the true function $\gamma(u) = 1.5[\sin(0.8u) + 1]$. All the u_i are independently generated from unif$[-6, 6]$.

To show the accuracy of our approach compared to INLA when the baseline hazard function is non-smooth, we did the following simulation:

- We simulate $\mathbf{N} = 400$ independent observations from a CoxPH Model with the following baseline hazard function:



- For simplicity, we assume the linear predictor is $\eta_i = \gamma(u_i) + \epsilon_i$, with the true function $\gamma(u) = 1.5[\sin(0.8u) + 1]$. All the u_i are independently generated from unif$[-6, 6]$.

- Among these 400 observations, we randomly censored 80 of them. The covariate u is discretized into 50 disjoint and equally-spaced bins.
- We then implemented the RW2 smoothing using both our method and INLA. In both cases, the variance parameter σ_u is set to have a PC prior such that $\mathbf{P}(\sigma_u > 2.5) = 0.5$ (Simpson et al., 2017).
- For the implementation of INLA, we used its default setting, which is to use a first-order random walk model for the baseline hazard.
- The results are summarized as the plot at the next page.

- Among these 400 observations, we randomly censored 80 of them. The covariate u is discretized into 50 disjoint and equally-spaced bins.
- We then implemented the RW2 smoothing using both our method and INLA. In both cases, the variance parameter σ_u is set to have a PC prior such that $\mathbf{P}(\sigma_u > 2.5) = 0.5$ (Simpson et al., 2017).
- For the implementation of INLA, we used its default setting, which is to use a first-order random walk model for the baseline hazard.
- The results are summarized as the plot at the next page.

- Among these 400 observations, we randomly censored 80 of them. The covariate u is discretized into 50 disjoint and equally-spaced bins.
- We then implemented the RW2 smoothing using both our method and INLA. In both cases, the variance parameter σ_u is set to have a PC prior such that $\mathbf{P}(\sigma_u > 2.5) = 0.5$ (Simpson et al., 2017).
- For the implementation of INLA, we used its default setting, which is to use a first-order random walk model for the baseline hazard.
- The results are summarized as the plot at the next page.

- Among these 400 observations, we randomly censored 80 of them. The covariate u is discretized into 50 disjoint and equally-spaced bins.
- We then implemented the RW2 smoothing using both our method and INLA. In both cases, the variance parameter σ_u is set to have a PC prior such that $\mathbf{P}(\sigma_u > 2.5) = 0.5$ (Simpson et al., 2017).
- For the implementation of INLA, we used its default setting, which is to use a first-order random walk model for the baseline hazard.
- The results are summarized as the plot at the next page.

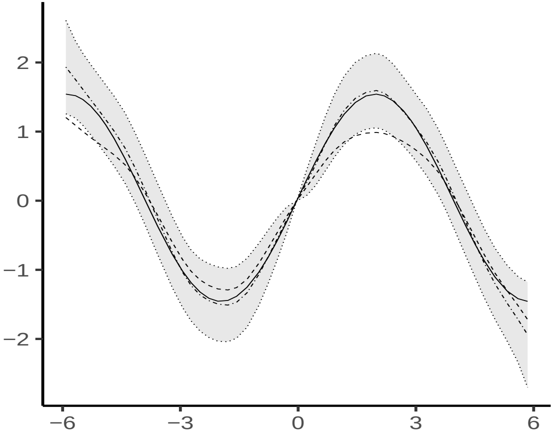


Figure: True risk function (—); posterior mean (- · -) and 95% credible interval (· · ·) using proposed method; posterior mean using INLA (- - -).

Martino et al. (2011) analyzed the Leukaemia data set using INLA (therefore, full-likelihood). The dataset contains information from 1043 independent adult leukaemia patients, with 16 percent of observations right-censored. Specifically, the main interest is to quantify the relationship between survival rate of leukaemia patients with the Townsend deprivation index (`tpi`) corresponding to the patient's location, conditional on the age of the patient, the count of white blood cells at diagnosis (`wbc`) and `sex` of the patient.

- The effects of `age`, `wbc` and `sex` were modelled linearly.Prior distributions $\beta \overset{iid}{\sim} \mathsf{N}(0, 0.001^{-1})$, were used for the linear effects.
- The `tpi` was discretized into 50 equally spaced bins and modelled as a semi-parametric (smoothing) effect. The semi-parametric effects $\Gamma = \{\gamma(\mathtt{tpi}_1), \cdots, \gamma(\mathtt{tpi}_{50})\}$ were modelled using the RW2 model with the reference constraint $\gamma(0) = 0$.
- The single variance parameter σ was given a PC prior such that $\mathsf{P}(\sigma > 2) = 0.5$.

- The effects of `age`, `wbc` and `sex` were modelled linearly.Prior distributions $\beta \overset{iid}{\sim} \mathrm{N}(0, 0.001^{-1})$, were used for the linear effects.
- The `tpi` was discretized into 50 equally spaced bins and modelled as a semi-parametric (smoothing) effect. The semi-parametric effects $\Gamma = \{\gamma(\mathtt{tpi}_1), \cdots, \gamma(\mathtt{tpi}_{50})\}$ were modelled using the RW2 model with the reference constraint $\gamma(0) = 0$.
- The single variance parameter σ was given a PC prior such that $\mathrm{P}(\sigma > 2) = 0.5$.

- The effects of `age`, `wbc` and `sex` were modelled linearly.Prior distributions $\beta \overset{iid}{\sim} \mathrm{N}(0, 0.001^{-1})$, were used for the linear effects.
- The `tpi` was discretized into 50 equally spaced bins and modelled as a semi-parametric (smoothing) effect. The semi-parametric effects $\Gamma = \{\gamma(\texttt{tpi}_1), \cdots, \gamma(\texttt{tpi}_{50})\}$ were modelled using the RW2 model with the reference constraint $\gamma(0) = 0$.
- The single variance parameter σ was given a PC prior such that $\mathrm{P}(\sigma > 2) = 0.5$.

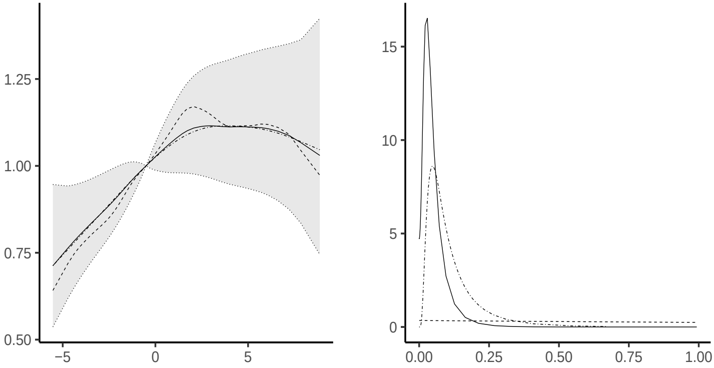


Figure: (a): posterior mean (—) and 95% credible interval (· · ·) using our method, posterior mean using INLA (- - -), and the result of fitting a GAM (- · -). (b): prior (- - -) and approximate posterior distribution for σ using our method (—) and INLA (- · -).

- We introduce an approximate Bayesian inference method for Cox Proportional Hazard model with **partial likelihood**.
- Because of the partial likelihood it used, the method does not have restriction on the form of baseline hazard function.
- The proposed method allows the inference of semi-parametric smoothing effect, that is is not sensitive to the number and placement of bins.
- The proposed method also allows observations to be correlated within subject (random intercept for subjects).
- Compared to traditional frequentist's method, Bayesian inference provided model-based estimation and uncertainty qualification.
- Due to the stable and fast optimization algorithm (quasi-Newton method), computations are **fast** for small to median sized data set.

- We introduce an approximate Bayesian inference method for Cox Proportional Hazard model with **partial likelihood**.
- Because of the partial likelihood it used, the method does not have restriction on the form of baseline hazard function.
- The proposed method allows the inference of semi-parametric smoothing effect, that is is not sensitive to the number and placement of bins.
- The proposed method also allows observations to be correlated within subject (random intercept for subjects).
- Compared to traditional frequentist's method, Bayesian inference provided model-based estimation and uncertainty qualification.
- Due to the stable and fast optimization algorithm (quasi-Newton method), computations are **fast** for small to median sized data set.

- We introduce an approximate Bayesian inference method for Cox Proportional Hazard model with **partial likelihood**.
- Because of the partial likelihood it used, the method does not have restriction on the form of baseline hazard function.
- The proposed method allows the inference of semi-parametric smoothing effect, that is is not sensitive to the number and placement of bins.
- The proposed method also allows observations to be correlated within subject (random intercept for subjects).
- Compared to traditional frequentist's method, Bayesian inference provided model-based estimation and uncertainty qualification.
- Due to the stable and fast optimization algorithm (quasi-Newton method), computations are **fast** for small to median sized data set.

- We introduce an approximate Bayesian inference method for Cox Proportional Hazard model with **partial likelihood**.
- Because of the partial likelihood it used, the method does not have restriction on the form of baseline hazard function.
- The proposed method allows the inference of semi-parametric smoothing effect, that is is not sensitive to the number and placement of bins.
- The proposed method also allows observations to be correlated within subject (random intercept for subjects).
- Compared to traditional frequentist's method, Bayesian inference provided model-based estimation and uncertainty qualification.
- Due to the stable and fast optimization algorithm (quasi-Newton method), computations are **fast** for small to median sized data set.

- We introduce an approximate Bayesian inference method for Cox Proportional Hazard model with **partial likelihood**.
- Because of the partial likelihood it used, the method does not have restriction on the form of baseline hazard function.
- The proposed method allows the inference of semi-parametric smoothing effect, that is is not sensitive to the number and placement of bins.
- The proposed method also allows observations to be correlated within subject (random intercept for subjects).
- Compared to traditional frequentist's method, Bayesian inference provided model-based estimation and uncertainty qualification.
- Due to the stable and fast optimization algorithm (quasi-Newton method), computations are **fast** for small to median sized data set.

- We introduce an approximate Bayesian inference method for Cox Proportional Hazard model with **partial likelihood**.
- Because of the partial likelihood it used, the method does not have restriction on the form of baseline hazard function.
- The proposed method allows the inference of semi-parametric smoothing effect, that is is not sensitive to the number and placement of bins.
- The proposed method also allows observations to be correlated within subject (random intercept for subjects).
- Compared to traditional frequentist's method, Bayesian inference provided model-based estimation and uncertainty qualification.
- Due to the stable and fast optimization algorithm (quasi-Newton method), computations are **fast** for small to median sized data set.

- More types of effects? For example: Spatial effect, temporal effect ... Need to include other covariance structure.
- More types of censoring? For example, besides traditional right-censoring, we can have: left-censoring, interval-censoring, left-truncations ... Need to modify the partial likelihood function.
- The method does not work efficient for data set that has very large size, as the computational cost grows as $O(n^2)$ (The fully dense Hessian matrix still needs to be evaluated at each maximum).

- More types of effects? For example: Spatial effect, temporal effect ... Need to include other covariance structure.
- More types of censoring? For example, besides traditional right-censoring, we can have: left-censoring, interval-censoring, left-truncations ... Need to modify the partial likelihood function.
- The method does not work efficient for data set that has very large size, as the computational cost grows as $O(n^2)$ (The fully dense Hessian matrix still needs to be evaluated at each maximum).

- More types of effects? For example: Spatial effect, temporal effect ... Need to include other covariance structure.
- More types of censoring? For example, besides traditional right-censoring, we can have: left-censoring, interval-censoring, left-truncations ... Need to modify the partial likelihood function.
- The method does not work efficient for data set that has very large size, as the computational cost grows as $O(n^2)$ (The fully dense Hessian matrix still needs to be evaluated at each maximum).

- More types of effects? For example: Spatial effect, temporal effect ... Need to include other covariance structure.
- More types of censoring? For example, besides traditional right-censoring, we can have: left-censoring, interval-censoring, left-truncations ... Need to modify the partial likelihood function.
- The method does not work efficient for data set that has very large size, as the computational cost grows as $O(n^2)$ (The fully dense Hessian matrix still needs to be evaluated at each maximum).

- More types of effects? For example: Spatial effect, temporal effect ... Need to include other covariance structure.
- More types of censoring? For example, besides traditional right-censoring, we can have: left-censoring, interval-censoring, left-truncations ... Need to modify the partial likelihood function.
- The method does not work efficient for data set that has very large size, as the computational cost grows as $O(n^2)$ (The fully dense Hessian matrix still needs to be evaluated at each maximum).

Braun, M. (2014). trustOptim: An R package for trust region optimization with sparse hessians. *Journal of Statistical Software 60*(4), 1–16.

Lindgren, F. and H. Rue (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics 35*(4), 691–700.

Martino, S., R. Akerkar, and H. Rue (2011). Approximate bayesian inference for survival models. *Scandinavian Journal of Statistics 38*(3), 514–528.

Simpson, D., H. Rue, T. G. Martins, A. Riebler, and S. H. Sørbye (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science 32*(1).

Stringer, A., P. Brown, and J. Stafford (2020). Approximate bayesian inference for case crossover models. *Biometrics*.

Tierney, L. and J. B. Kadane (1986). Accurate approximations to posterior moments and marginal densities. *Journal of the*