Bayesian inference using Catalytic prior distributions for Cox regression models

Weihao Li

Joint work with Dongming Huang

Department of Statistics, NUS

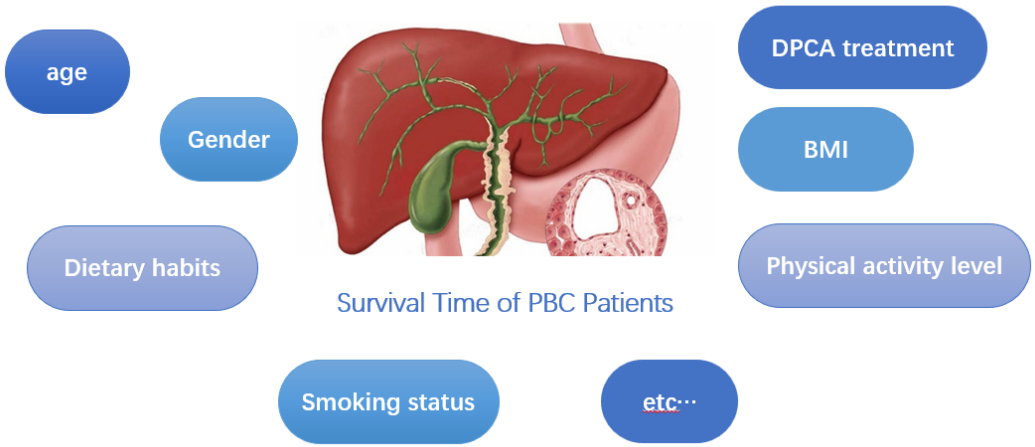
Dec 6, 2023

Motivation

- Survival analysis, often used in medical research, epidemiology.

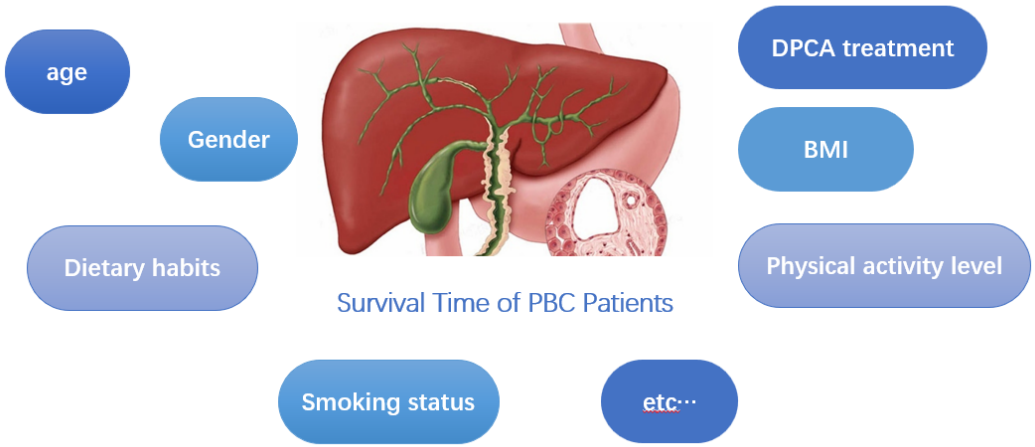
Motivation

- Survival analysis, often used in medical research, epidemiology.



Motivation

- Survival analysis, often used in medical research, epidemiology.



- Large dimension & insufficient sample.

Motivation

- Cox regression model: popular way in survival analysis

Motivation

- Cox regression model: popular way in survival analysis

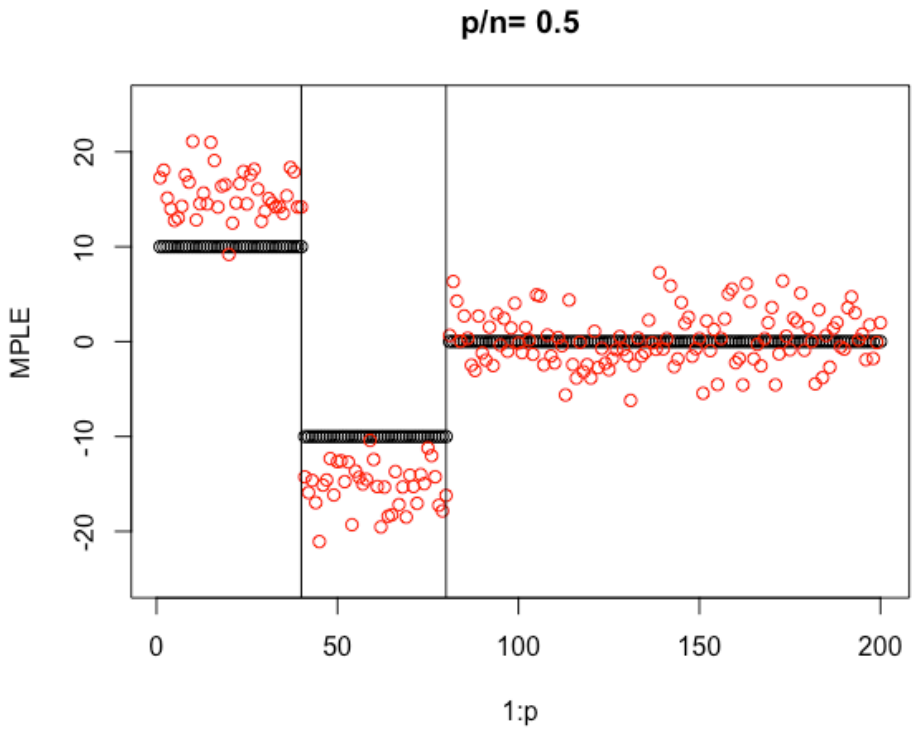
```
library(survival)
coxph(Surv(time, status) ~ X)
```

Motivation

- Cox regression model: popular way in survival analysis

```
library(survival)
coxph(Surv(time, status) ~ X)
```

- [Zhang et al., 2022] $p = 200, n = 400$



Introduction of Cox regression model

- Observed data: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n$, $\boldsymbol{X}_i \in \mathbb{R}^p$ is the feature vector, δ_i is the censoring indicator, and $Y_i = \min(T_i, C_i)$, where T_i is the survival time for an uncensored subject and C_i is the censoring time for a censored subject.

Introduction of Cox regression model

- Observed data: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n$, $\boldsymbol{X}_i \in \mathbb{R}^p$ is the feature vector, δ_i is the censoring indicator, and $Y_i = \min(T_i, C_i)$, where T_i is the survival time for an uncensored subject and C_i is the censoring time for a censored subject.
- Cox regression model: semi-parametric

$$L(\beta, h_0 \mid \boldsymbol{D}) = \prod_{i=1}^{n} \left\{\exp\left(\boldsymbol{X}_i'\beta\right) h_0(Y_i)\right\}^{\delta_i} \exp\left\{-\exp\left(\boldsymbol{X}_i'\beta\right) H_0(Y_i)\right\}$$

  - $h_0(t)$: baseline hazard function, unknown nuisance parameter

Introduction of Cox regression model

- Observed data: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n$, $\boldsymbol{X}_i \in \mathbb{R}^p$ is the feature vector, δ_i is the censoring indicator, and $Y_i = \min(T_i, C_i)$, where T_i is the survival time for an uncensored subject and C_i is the censoring time for a censored subject.
- Cox regression model: semi-parametric

$$L(\beta, h_0 \mid \boldsymbol{D}) = \prod_{i=1}^{n} \left\{ \exp\left(\boldsymbol{X}_i'\beta\right) h_0\left(Y_i\right) \right\}^{\delta_i} \exp\left\{ -\exp\left(\boldsymbol{X}_i'\beta\right) H_0\left(Y_i\right) \right\}$$

  - $h_0(t)$: baseline hazard function, unknown nuisance parameter
- Partial likelihood [Cox, 1975]: MPLE maximizes

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp\left(\boldsymbol{X}_i'\beta\right)}{\sum_{l \in \mathbf{R}_i} \exp\left(\boldsymbol{X}_l'\beta\right)} \right\}^{\delta_i}$$

where $\mathbf{R}_i := \{j : Y_j \geq Y_i\}$ is the collection of index for those are risky at time Y_i.

Recap the problem of classical method

- When dimension is large or insufficient sample, MPLE suffers
  1. highly variable
  2. not exist
  3. large bias

Recap the problem of classical method

- When dimension is large or insufficient sample, MPLE suffers
  1. highly variable
  2. not exist
  3. large bias
- Similar problems for MLE

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \left(\prod_{i}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}) \right)$$

  - large variability for large p
  - MLE may not exist, e.g.: Logistic regression
  - large bias

Related methods to address issues

- Existing strategy for MLE:
  - [Huang et al., 2020] proposed class of automatic priors called **catalytic prior** for GLM to provide stable estimation in high dimensional model.

Related methods to address issues

- Existing strategy for MLE:
  - [Huang et al., 2020] proposed class of automatic priors called **catalytic prior** for GLM to provide stable estimation in high dimensional model.
  - Increase 'sample size' by generating synthetic data $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$.

$$Obs : \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n \qquad Syn : \{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$$

Related methods to address issues

- Existing strategy for MLE:
  - [Huang et al., 2020] proposed class of automatic priors called **catalytic prior** for GLM to provide stable estimation in high dimensional model.
  - Increase 'sample size' by generating synthetic data $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$.

$$Obs : \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n \qquad Syn : \{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$$

  - Catalytic prior based on synthetic data

$$\pi_{cat,M}(\boldsymbol{\theta} \mid \tau) \propto \left(\prod_{i=1}^{M} f\left(Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\theta}\right) \right)^{\tau/M}$$

Related methods to address issues

- Existing strategy for MLE:
  - [Huang et al., 2020] proposed class of automatic priors called **catalytic prior** for GLM to provide stable estimation in high dimensional model.
  - Increase 'sample size' by generating synthetic data $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$.

$$Obs : \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n \qquad Syn : \{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$$

  - Catalytic prior based on synthetic data

$$\pi_{cat,M}(\boldsymbol{\theta} \mid \tau) \propto \left(\prod_{i=1}^{M} f\left(Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\theta}\right) \right)^{\tau/M}$$

  - Posterior distribution:

$$\pi(\boldsymbol{\theta} \mid \{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n) \propto \left(\prod_{i}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}) \right) \pi_{cat,M}(\boldsymbol{\theta} \mid \tau)$$

$$\propto \exp\left\{ \sum_{i=1}^{n} \log\left(f\left(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}\right)\right) + \frac{\tau}{M} \sum_{i=1}^{M} \log\left(f\left(Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\theta}\right)\right) \right\}$$

Stable estimation for simpler model

- Experts focus on several important covariates.
  E.g.: thyroid-related diseases–gender, emotional state, age, etc.

[1] Other type of simpler models are possible

Stable estimation for simpler model

- Experts focus on several important covariates.
  E.g.: thyroid-related diseases–gender, emotional state, age, etc.
  Only fit simpler model with subset of covariates $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$ [1].
  $\rightarrow \hat{\psi} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \rightarrow$ predictive distribution $g_* (y \mid \boldsymbol{x})$

[1] Other type of simpler models are possible

Stable estimation for simpler model

- Experts focus on several important covariates.
  E.g.: thyroid-related diseases–gender, emotional state, age, etc.
  Only fit simpler model with subset of covariates $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$ [1].
  $\rightarrow \hat{\psi} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \rightarrow$ predictive distribution $g_* (y \mid \boldsymbol{x})$
- Create fake data to "increase" sample size. M synthetic data points $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$ can be generated according to following strategy:

$$\boldsymbol{X}_i^* \overset{i.i.d.}{\sim} Q(\boldsymbol{x}), \quad Y_i^* \mid \boldsymbol{X}_i^* \sim g_* (y \mid \boldsymbol{X}_i^*)$$

[1] Other type of simpler models are possible

Stable estimation for simpler model

- Experts focus on several important covariates.
  E.g.: thyroid-related diseases–gender, emotional state, age, etc.
  Only fit simpler model with subset of covariates $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3\}$ [1].
  $\rightarrow \hat{\psi} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \rightarrow$ predictive distribution $g_* (y \mid \boldsymbol{x})$
- Create fake data to "increase" sample size. M synthetic data points $\{(Y_i^*, \boldsymbol{X}_i^*)\}_{i=1}^M$ can be generated according to following strategy:

$$\boldsymbol{X}_i^* \overset{i.i.d.}{\sim} Q(\boldsymbol{x}), \quad Y_i^* \mid \boldsymbol{X}_i^* \sim g_* (y \mid \boldsymbol{X}_i^*)$$

- Now we are able to define catalytic prior for $\boldsymbol{\beta}$

$$\pi_{cat,M}(\boldsymbol{\theta} \mid \tau) \propto \left(\prod_{i=1}^{M} f (Y_i^* \mid \boldsymbol{X}_i^*, \boldsymbol{\theta}) \right)^{\tau/M}$$

[1] Other type of simpler models are possible

Generation of synthetic survival data

- Synthetic covariates: $\boldsymbol{X}_i^* \overset{i.i.d.}{\sim} Q(\boldsymbol{x})$
- Synthetic survival time:
  1. Simpler predictive model $g(y \mid \boldsymbol{X}, \boldsymbol{\delta}, \boldsymbol{\psi})$ with $\boldsymbol{\psi}$ can be stably fitted from the observed data $\{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n$, denote the predictive distribution as $g_*(y \mid \boldsymbol{x}, \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n)$
  2. $Y_i^* \mid \boldsymbol{X}_i^* \sim g_*\left(y \mid \boldsymbol{X}_i^*, \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n\right)$

Generation of synthetic survival data

- Synthetic covariates: $\boldsymbol{X}_i^* \overset{i.i.d.}{\sim} Q(\boldsymbol{x})$
- Synthetic survival time:
  1. Simpler predictive model $g(y \mid \boldsymbol{X}, \boldsymbol{\delta}, \boldsymbol{\psi})$ with $\boldsymbol{\psi}$ can be stably fitted from the observed data $\{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n$, denote the predictive distribution as $g_*(y \mid \boldsymbol{x}, \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n)$
  2. $Y_i^* \mid \boldsymbol{X}_i^* \sim g_*\left(y \mid \boldsymbol{X}_i^*, \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n\right)$

  Remark: all synthetic survival subjects are uncensored.

Generation of synthetic survival data

- Synthetic covariates: $\boldsymbol{X}_i^* \overset{i.i.d.}{\sim} Q(\boldsymbol{x})$
- Synthetic survival time:
  1. Simpler predictive model $g(y \mid \boldsymbol{X}, \boldsymbol{\delta}, \boldsymbol{\psi})$ with $\boldsymbol{\psi}$ can be stably fitted from the observed data $\{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n$, denote the predictive distribution as $g_*(y \mid \boldsymbol{x}, \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n)$
  2. $Y_i^* \mid \boldsymbol{X}_i^* \sim g_* \left(y \mid \boldsymbol{X}_i^*, \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n\right)$

  Remark: all synthetic survival subjects are uncensored.

Example:

1. Simpler model:

$$L\left(\psi \mid \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n\right) = \prod_{i=1}^n \psi^{\delta_i} \exp\{-\psi Y_i\}, \quad \psi > 0.$$

2. $Y_i^* \mid \boldsymbol{X}_i^* \sim \text{Exp}(\hat{\psi})$

Natural extension: Catalytic prior on (β, h_0)

$$\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}$$

$$\pi_{cat,both}(\beta, h_0) \propto L(\beta, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

Natural extension: Catalytic prior on (β, h_0)

$$\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M$$

$$\pi_{cat,both}(\beta, h_0) \propto L(\beta, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

- Posterior distribution:

$$\pi_{post,both}(\beta, h_0) \propto L(\beta, h_0(\cdot) \mid \boldsymbol{D}) \cdot L(\beta, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

Natural extension: Catalytic prior on (β, h_0)

$$\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}$$

$$\pi_{cat,both}(\beta, h_0) \propto L(\beta, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

- Posterior distribution:

$$\pi_{post,both}(\beta, h_0) \propto L(\beta, h_0(\cdot) \mid \boldsymbol{D}) \cdot L(\beta, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

Conceptually valid but not able to be incorporated into standard Bayesian Cox model.
Reason: β and h_0 are prior-dependent

Natural extension: Catalytic prior on (β, h_0)

$$\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M$$

$$\pi_{cat,both}(\beta, h_0) \propto L(\beta, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

- Posterior distribution:

$$\pi_{post,both}(\beta, h_0) \propto L(\beta, h_0(\cdot) \mid \boldsymbol{D}) \cdot L(\beta, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

  Conceptually valid but not able to be incorporated into standard Bayesian Cox model.
  Reason: β and h_0 are prior-dependent
- Efficient point estimation for β: profile $h_0(\cdot)$ out [Murphy and Van der Vaart, 2000]

Weighted mixture estimator

Algorithm: Compute **W**eighted **M**ixture estimator in R
Require: **R** package `survival`
Input: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}, \tau > 0$

Weighted mixture estimator

Algorithm: Compute **W**eighted **M**ixture estimator in R

Require: **R** package `survival`

Input: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}, \tau > 0$

1. Combine data:
   $\tilde{\mathbb{X}} \leftarrow \text{rbind}\,(\mathbb{X}, \mathbb{X}^*)$

   $\tilde{\boldsymbol{Y}} \leftarrow c\,(\boldsymbol{Y}, \boldsymbol{Y}^*)$

   $\tilde{\delta} \leftarrow c(\delta, \mathbf{1})$

Weighted mixture estimator

Algorithm: Compute **W**eighted **M**ixture estimator in R

Require: **R** package `survival`

Input: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}, \tau > 0$

1. Combine data:
   $\tilde{\mathbb{X}} \leftarrow \text{rbind}\,(\mathbb{X}, \mathbb{X}^*)$

   $\tilde{\boldsymbol{Y}} \leftarrow c\,(\boldsymbol{Y}, \boldsymbol{Y}^*)$

   $\tilde{\delta} \leftarrow c(\delta, \mathbf{1})$
2. Compute weight vector:
   $\tilde{\mathbf{w}} \leftarrow c(\text{rep}(1, n), \text{rep}(\frac{\tau}{M}, M))$

Weighted mixture estimator

Algorithm: Compute **W**eighted **M**ixture estimator in R
Require: **R** package `survival`
Input: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}, \tau > 0$

1. Combine data:
   $\tilde{\mathbb{X}} \leftarrow \text{rbind}\,(\mathbb{X}, \mathbb{X}^*)$
   $\tilde{\boldsymbol{Y}} \leftarrow c\,(\boldsymbol{Y}, \boldsymbol{Y}^*)$
   $\tilde{\delta} \leftarrow c(\delta, \mathbf{1})$
2. Compute weight vector:
   $\tilde{\mathbf{w}} \leftarrow c(\text{rep}(1, n), \text{rep}(\frac{\tau}{M}, M))$
3. Fit proportional hazards regression model with combined data and weight vector:
   $\text{fit} \leftarrow \text{coxph}(\text{Surv}(\tilde{\boldsymbol{Y}}, \tilde{\delta}) \sim \tilde{\mathbb{X}}, \text{weights} = \tilde{\mathbf{w}})$

Weighted mixture estimator

Algorithm: Compute **W**eighted **M**ixture estimator in R

Require: **R** package `survival`

Input: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}, \tau > 0$

1. Combine data:
   $\tilde{\mathbb{X}} \leftarrow \text{rbind}\,(\mathbb{X}, \mathbb{X}^*)$
   $\tilde{\boldsymbol{Y}} \leftarrow c\,(\boldsymbol{Y}, \boldsymbol{Y}^*)$
   $\tilde{\boldsymbol{\delta}} \leftarrow c(\boldsymbol{\delta}, \mathbf{1})$
2. Compute weight vector:
   $\tilde{\mathbf{w}} \leftarrow c(\text{rep}(1, n), \text{rep}(\frac{\tau}{M}, M))$
3. Fit proportional hazards regression model with combined data and weight vector:
   $\text{fit} \leftarrow \text{coxph}(\text{Surv}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{\delta}}) \sim \tilde{\mathbb{X}}, \text{weights} = \tilde{\mathbf{w}})$
4. $\hat{\boldsymbol{\beta}}_{WM,\tau} = \text{coef}(\text{fit})$

Weighted mixture estimator

Algorithm: Compute **W**eighted **M**ixture estimator in R

Require: **R** package `survival`

Input: $\boldsymbol{D} = \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^{n}, \boldsymbol{D}^* = \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^{M}, \tau > 0$

1. Combine data:
   $\tilde{\mathbb{X}} \leftarrow \text{rbind}\,(\mathbb{X}, \mathbb{X}^*)$
   $\tilde{\boldsymbol{Y}} \leftarrow c\,(\boldsymbol{Y}, \boldsymbol{Y}^*)$
   $\tilde{\boldsymbol{\delta}} \leftarrow c(\boldsymbol{\delta}, \mathbf{1})$
2. Compute weight vector:
   $\tilde{\mathbf{w}} \leftarrow c(\text{rep}(1, n), \text{rep}(\frac{\tau}{M}, M))$
3. Fit proportional hazards regression model with combined data and weight vector:
   $\text{fit} \leftarrow \text{coxph}(\text{Surv}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{\delta}}) \sim \tilde{\mathbb{X}}, \text{weights} = \tilde{\mathbf{w}})$
4. $\hat{\boldsymbol{\beta}}_{WM,\tau} = \text{coef}(\text{fit})$

Property:

- $\hat{\boldsymbol{\beta}}_{WM,\tau}$ is consistent when p is fixed and $\tau = o(n)$.

Catalytic prior for Cox model on β

- Previous catalytic prior for GLM rely on known likelihood function.
- Issue: $h_0(t)$ is unknown nuisance parameter in Cox model

Catalytic prior for Cox model on β

- Previous catalytic prior for GLM rely on known likelihood function.
- Issue: $h_0(t)$ is unknown nuisance parameter in Cox model

$$\pi_{cat,both}(\boldsymbol{\beta}, h_0) \propto L(\boldsymbol{\beta}, h_0(\cdot) \mid \boldsymbol{D}^*)^{\frac{\tau}{M}}$$

$$= \left[\prod_{i=1}^{M} \left\{ \exp\left(\boldsymbol{X}_i^{*\prime}\boldsymbol{\beta}\right) h_0(Y_i^*) \right\} \exp\left\{ -\exp\left(\boldsymbol{X}_i^{*\prime}\boldsymbol{\beta}\right) \int_0^{Y_i^*} h_0(s)ds \right\} \right]^{\tau/M}$$

Catalytic prior for Cox model on β

- Previous catalytic prior for GLM rely on known likelihood function.
- Issue: $h_0(t)$ is unknown nuisance parameter in Cox model

$$\pi_{cox,cat}(\beta \mid \tau) \propto L\left(\beta, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M}$$

$$= \left[\prod_{i=1}^M \left\{\exp\left(\boldsymbol{X}_i^{*\prime}\beta\right) h_0^+\right\} \exp\left\{-\exp\left(\boldsymbol{X}_i^{*\prime}\beta\right) Y_i^* h_0^+\right\}\right]^{\tau/M}$$

Catalytic prior for Cox model on β

- Previous catalytic prior for GLM rely on known likelihood function.
- Issue: $h_0(t)$ is unknown nuisance parameter in Cox model

$$\pi_{cox,cat}(\beta \mid \tau) \propto L\left(\beta, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M}$$
$$= \left[\prod_{i=1}^M \left\{\exp\left(\boldsymbol{X}_i^{*\prime}\beta\right) h_0^+\right\} \exp\left\{-\exp\left(\boldsymbol{X}_i^{*\prime}\beta\right) Y_i^* h_0^+\right\}\right]^{\tau/M}$$

- User-specific surrogate baseline hazard constant $h_0^+ > 0$
  1. User-Defined constant: constant hazard, hazard with domain knowledge
  2. Computed in a data driven way.

Catalytic prior for Cox model on β

- Previous catalytic prior for GLM rely on known likelihood function.
- Issue: $h_0(t)$ is unknown nuisance parameter in Cox model

$$\pi_{cox,cat}(\beta \mid \tau) \propto L\left(\beta, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M}$$
$$= \left[\prod_{i=1}^M \left\{\exp\left(\boldsymbol{X}_i^{*\prime}\beta\right) h_0^+\right\} \exp\left\{-\exp\left(\boldsymbol{X}_i^{*\prime}\beta\right) Y_i^* h_0^+\right\}\right]^{\tau/M}$$

- User-specific surrogate baseline hazard constant $h_0^+ > 0$
  1. User-Defined constant: constant hazard, hazard with domain knowledge
  2. Computed in a data driven way.

  h_0^+ acts merely as a surrogate for the nuisance component to facilitate the construction of our catalytic prior. It does not need to be correctly specified or unbiased.

- Example on h_0^+ :
  1. $h_0^+ = \hat{\psi}$

$$\hat{\psi} = \arg\max_{\psi} \left(\prod_{i=1}^{n} \psi^{\delta_i} \exp\{-\psi Y_i\} \right), \quad \psi > 0.$$

  2. $h_0^+ = 1/\bar{Y}$

Bayesian Cox model with catalytic prior

$$\pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) = \quad L\left(\boldsymbol{\beta}, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M} \tag{1}$$

Procedure:

- Standard Bayesian Cox model with $L(\boldsymbol{\beta}, h_0)$ as likelihood and default prior for parameter $\pi_0(\boldsymbol{\beta}) \cdot \pi(h_0)$

$$\pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) \propto L(\boldsymbol{\beta}, h_0) \cdot \pi_0(\boldsymbol{\beta}) \cdot \pi(h_0)$$

Bayesian Cox model with catalytic prior

$$\pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) = \quad L\left(\boldsymbol{\beta}, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M} \tag{1}$$

Procedure:

- Standard Bayesian Cox model with $L(\boldsymbol{\beta}, h_0)$ as likelihood and default prior for parameter $\pi_0(\boldsymbol{\beta}) \cdot \pi(h_0)$

$$\pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) \propto L(\boldsymbol{\beta}, h_0) \cdot \pi_0(\boldsymbol{\beta}) \cdot \pi(h_0)$$

- Replace default prior $\pi_0(\boldsymbol{\beta})$ on $\boldsymbol{\beta}$ by catalytic prior Eq.(1).
- Run MCMC to collect posterior samples.

Bayesian Cox model with catalytic prior

$$\pi_{cox,cat}(\beta \mid \tau) = \quad L\left(\beta, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M} \tag{1}$$

Procedure:

- Standard Bayesian Cox model with $L(\beta, h_0)$ as likelihood and default prior for parameter $\pi_0(\beta) \cdot \pi(h_0)$

$$\pi_{post}(\beta, h_0 \mid \boldsymbol{D}) \propto L(\beta, h_0) \cdot \pi_0(\beta) \cdot \pi(h_0)$$

- Replace default prior $\pi_0(\beta)$ on β by catalytic prior Eq.(1).
- Run MCMC to collect posterior samples.

Property:

- $\pi_{cox,cat}(\beta \mid \tau)$ is proper under mild assumption.

Approximate posterior mode

$$\pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) = L(\boldsymbol{\beta}, h_0) \cdot \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) \cdot \pi(h_0)$$

Approximate posterior mode

$$\pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) = L(\boldsymbol{\beta}, h_0) \cdot \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) \cdot \pi(h_0)$$

$$\arg\max_{\boldsymbol{\beta}} \ \pi_{margin}(\boldsymbol{\beta} \mid \boldsymbol{D}) = \arg\max_{\boldsymbol{\beta}} \ \int \pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) dh_0$$

Approximate posterior mode

$$\pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) = L(\boldsymbol{\beta}, h_0) \cdot \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) \cdot \pi(h_0)$$

$$\arg\max_{\boldsymbol{\beta}} \ \pi_{margin}(\boldsymbol{\beta} \mid \boldsymbol{D}) = \arg\max_{\boldsymbol{\beta}} \ \int \pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) dh_0$$

Logic: from [Sinha, 2003], when $\pi(h_0)$ is diffuse,

$$\pi_{margin}(\boldsymbol{\beta} \mid \boldsymbol{D}) \approx PL(\boldsymbol{\beta}) \cdot \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau)$$

Approximate posterior mode

$$\pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) = L(\boldsymbol{\beta}, h_0) \cdot \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) \cdot \pi(h_0)$$

$$\arg\max_{\boldsymbol{\beta}} \ \pi_{margin}(\boldsymbol{\beta} \mid \boldsymbol{D}) = \arg\max_{\boldsymbol{\beta}} \int \pi_{post}(\boldsymbol{\beta}, h_0 \mid \boldsymbol{D}) dh_0$$

Logic: from [Sinha, 2003], when $\pi(h_0)$ is diffuse,

$$\pi_{margin}(\boldsymbol{\beta} \mid \boldsymbol{D}) \approx PL(\boldsymbol{\beta}) \cdot \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau)$$

Point estimation (**C**atalytic **R**egularized estimator)

$$\hat{\boldsymbol{\beta}}_{CR,\tau} = \arg\max_{\boldsymbol{\beta}} \{\log PL(\boldsymbol{\beta}) + \log \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau)\}, \tag{2}$$

Approximate posterior mode

$$\pi_{post}(\beta, h_0 \mid \boldsymbol{D}) = L(\beta, h_0) \cdot \pi_{cox,cat}(\beta \mid \tau) \cdot \pi(h_0)$$

$$\arg\max_{\beta} \ \pi_{margin}(\beta \mid \boldsymbol{D}) = \arg\max_{\beta} \int \pi_{post}(\beta, h_0 \mid \boldsymbol{D}) dh_0$$

Logic: from [Sinha, 2003], when $\pi(h_0)$ is diffuse,

$$\pi_{margin}(\beta \mid \boldsymbol{D}) \approx PL(\beta) \cdot \pi_{cox,cat}(\beta \mid \tau)$$

Point estimation (**C**atalytic **R**egularized estimator)

$$\hat{\beta}_{CR,\tau} = \arg\max_{\beta} \{\log PL(\beta) + \log \pi_{cox,cat}(\beta \mid \tau)\}, \tag{2}$$

Property:

- $\hat{\beta}_{CR,\tau}$ is consistent when p is fixed and $\tau = o(n)$.

Short summary

Method:

- Full Bayesian Procedure

$$\pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) \quad = \quad L\left(\boldsymbol{\beta}, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M}$$

- **C**atalytic **R**egularized estimator

$$\hat{\boldsymbol{\beta}}_{CR,\tau} = \arg\max_{\boldsymbol{\beta}} \{\log PL(\boldsymbol{\beta}) + \log \pi_{cox,cat}(\boldsymbol{\beta} \mid \tau)\}$$

- **W**eighted **M**ixture estimator

$$\hat{\boldsymbol{\beta}}_{WM,\tau}$$

Theory:

- $\pi_{cox,cat}(\boldsymbol{\beta} \mid \tau)$ is proper
- $\hat{\boldsymbol{\beta}}_{CR,\tau}$ and $\hat{\boldsymbol{\beta}}_{WM,\tau}$ are consistent.

Numerical studies

Observed data:

- Covariates:

$$\boldsymbol{X}_{i,j} \sim \begin{cases} \text{Bernoulli}(0.1), & j = 1 \\ \chi^2_1, & j = 2 \\ \chi^2_4, & j = 3 \\ N(0,1), & 4 \leq j \leq p \end{cases}$$

Numerical studies

Observed data:

- Covariates:

$$\boldsymbol{X}_{i,j} \sim \begin{cases} \text{Bernoulli}(0.1), & j = 1 \\ \chi^2_1, & j = 2 \\ \chi^2_4, & j = 3 \\ N(0,1), & 4 \leq j \leq p \end{cases}$$

- $n = 100$, 20% censored subjects.
- The true regression coefficient vector is set to be $\boldsymbol{\beta}_0 = (4, -4, 3, -3, \mathbf{1}_{p-4})/\sqrt{p}$.
- T_i independently from an exponential distribution with a rate parameter of $0.5 \exp(\boldsymbol{X}_i^\top \boldsymbol{\beta}_0)$.

Numerical studies

Synthetic data:

- M=1000
- Synthetic covariates: independent resampling from original covariates

Numerical studies

Synthetic data:

- M=1000
- Synthetic covariates: independent resampling from original covariates
  Modification:
  - half of the sampled X_1^* will be replaced by i.i.d. random variables draw from Bernoulli($p = 0.5$)—*flattening*
  - continuous covariate ($j \geq 2$), half of sampled X_j^* will be replaced by i.i.d random variables draw from a normal distribution with median and interquartile range matching to those of the observed covariates.

Numerical studies

Synthetic data:

- M=1000
- Synthetic covariates: independent resampling from original covariates
  Modification:
  - half of the sampled X_1^* will be replaced by i.i.d. random variables draw from Bernoulli($p = 0.5$)—*flattening*
  - continuous covariate ($j \geq 2$), half of sampled X_j^* will be replaced by i.i.d random variables draw from a normal distribution with median and interquartile range matching to those of the observed covariates.
- Synthetic survival time: $Y_i^* \sim \text{Exp}(\hat{\psi})$

$$L\left(\psi \mid \{(\boldsymbol{X}_i, Y_i, \delta_i)\}_{i=1}^n\right) = \prod_{i=1}^n \psi^{\delta_i} \exp\left\{-\psi Y_i\right\}, \quad \psi > 0.$$

Numerical studies

Our methods: $h_0^+ = \hat{\psi}$

$$\pi_{cox,cat}(\boldsymbol{\beta} \mid \tau) \quad = \quad L\left(\boldsymbol{\beta}, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M}$$

1. CRE: catalytic regularized estimator $\hat{\boldsymbol{\beta}}_{CR,\tau}$
2. WME: weighted mixture estimator $\hat{\boldsymbol{\beta}}_{WM,\tau}$
3. CPM: posterior mean of MCMC sampler based on catalytic prior

Alternative methods:

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \sum_{i=1}^n \delta_i \left(\boldsymbol{X}_i{}^\top \boldsymbol{\beta} - \log \sum_{j\in\mathcal{R}_i} \theta_j \right) - \lambda f(\boldsymbol{\beta}) \right\}$$

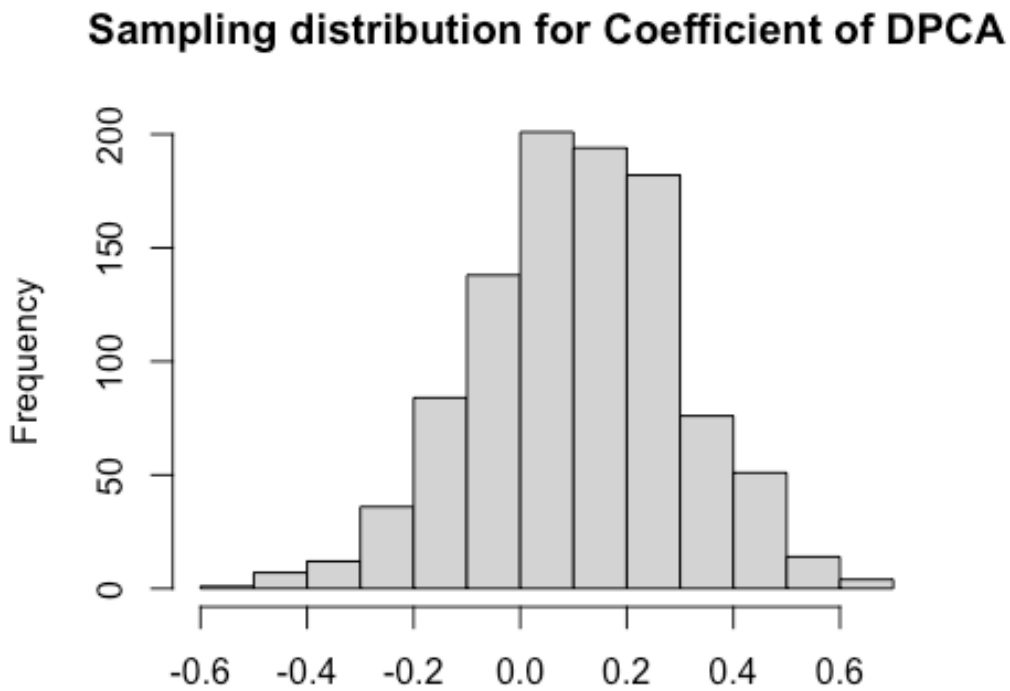
1. $f(\boldsymbol{\beta}) = 0 \Rightarrow$ MPLE: maximum partial likelihood estimator
2. $f(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 \Rightarrow$ Lasso
3. $f(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 \Rightarrow$ Ridge

$$Deviance(\boldsymbol{\beta}_0, \hat{\boldsymbol{\beta}}) = \left\{ \ell^{(\text{test})}(\boldsymbol{\beta}_0) - \ell^{(\text{test})}(\hat{\boldsymbol{\beta}}) \right\}$$

p	Methods	$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$	Predictive deviance
20	MPLE	0.95(0.06)	19.73(1.32)
	CRE (CV)	0.63(0.04)	12.87(0.80)
	CRE ($\tau = p$)	0.86(0.02)	19.53(0.70)
	WME (CV)	**0.51**(0.02)	**12.44**(0.75)
	CPM (CV)	0.79(0.04)	18.82(0.67)
	Ridge (CV)	0.58(0.03)	13.07(0.63)
	Lasso (CV)	0.75(0.04)	13.42(0.63)
40	MPLE	4.25(0.21)	103.38(4.99)
	CRE (CV)	0.82(0.02)	24.09(0.84)
	CRE ($\tau = p$)	0.88(0.01)	25.63(0.77)
	WME (CV)	**0.76**(0.02)	**23.37**(0.82)
	CPM (CV)	0.84(0.02)	24.63(0.76)
	Ridge (CV)	0.94(0.03)	24.53(0.87)
	Lasso (CV)	1.19(0.03)	24.06(0.92)

Real data: PBC dataset

- Data: $p = 18$, DPCA treatment + 17 covariates
- Full Bayesian analysis with $\tau = p$



95% Credible interval: [-0.28,0.47]. Ineffectiveness of DPCA treatment has been reported [Locke III et al., 1996].

Summary

- Propose Cox catalytic prior on $\boldsymbol{\beta}$ and its corresponding approximate posterior mode.
- Derive weighted mixture estimator Using synthetic data to derive.
- Establish properness of prior and consistency of point estimation.
- Show proposed methods outperform classical MPLE and comparable with existing regularization method.

Summary

- Propose Cox catalytic prior on $\boldsymbol{\beta}$ and its corresponding approximate posterior mode.
- Derive weighted mixture estimator Using synthetic data to derive.
- Establish properness of prior and consistency of point estimation.
- Show proposed methods outperform classical MPLE and comparable with existing regularization method.

Future direction:

- Other statistical inference that involve partial likelihood or complex nuisance parameters.
- Other semi-parametric model.

http://arxiv.org/abs/2312.01411

Thank you!

Appendix: Prior on τ

If a fully Bayesian perspective on τ is adopted, we can impose a joint prior on (τ, β) as follows. Given any two positive scalar hyperparameters α and γ, we define a joint catalytic prior for $(\tau, \boldsymbol{\beta})$ as

$$\pi_{\alpha,\gamma}(\tau, \boldsymbol{\beta}) \propto \Gamma_{\alpha,\gamma}(\tau) \cdot L\left(\boldsymbol{\beta}, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right)^{\tau/M} \tag{3}$$

where $\Gamma_{\alpha,\gamma}(\tau)$ is a function defined as

$$\Gamma_{\alpha,\gamma}(\tau) = \tau^{p+\alpha-1} e^{-\tau\left(\kappa+\gamma^{-1}\right)} \tag{4}$$

and

$$\kappa := \sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{M} \log L\left(\boldsymbol{\beta}, h_0^+ \mid \{(\boldsymbol{X}_i^*, Y_i^*)\}_{i=1}^M\right) \tag{5}$$

Appendix: Ridge and Lasso

The ridge estimator is defined as follows:

$$\hat{\boldsymbol{\beta}}_{\lambda} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \delta_i \left(\boldsymbol{X}_i{}^{\top}\boldsymbol{\beta} - \log \sum_{j\in\mathcal{R}_i} \theta_j \right) - \lambda f(\boldsymbol{\beta}) \right\} \tag{6}$$

where $\theta_j = \exp(\boldsymbol{X}_j^{\top}\boldsymbol{\beta})$ and the penalty term $f(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|^2$, assuming that each entry of the covariates $\boldsymbol{X}_i$'s has been standardized to have zero mean and unit variance. The Lasso estimator is defined in a similar way as in (6) but with $f(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. Both the ridge and the Lasso estimates can be computed using the `glmnet` package in **R** [Simon et al., 2011].

Reference I


Cox, D. R. (1975).
Partial likelihood.
Biometrika, 62(2):269–276.


Huang, D., Stein, N., Rubin, D. B., and Kou, S. C. (2020).
Catalytic prior distributions with application to generalized linear models.
Proc. Natl. Acad. Sci. U.S.A., 117(22):12004–12010.


Locke III, G. R., Therneau, T. M., Ludwig, J., Dickson, E. R., and Lindor, K. D. (1996).
Time course of histological progression in primary biliary cirrhosis.
Hepatology, 23(1):52–56.


Murphy, S. A. and Van der Vaart, A. W. (2000).
On profile likelihood.
Journal of the American Statistical Association, 95(450):449–465.

Reference II


Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011).
Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.
Journal of Statistical Software, 39(5):1–13.


Sinha, D. (2003).
A Bayesian justification of Cox's partial likelihood.
Biometrika, 90(3):629–641.


Zhang, X., Zhou, H., and Ye, H. (2022).
A modern theory for high-dimensional Cox regression models.
arXiv preprint arXiv:2204.01161.