Tutorial: Bayesian variable selection for survival data

Francisco Javier Rubio

University College London
Department of Statistical Science

August 2023

Overview

GitHub repository

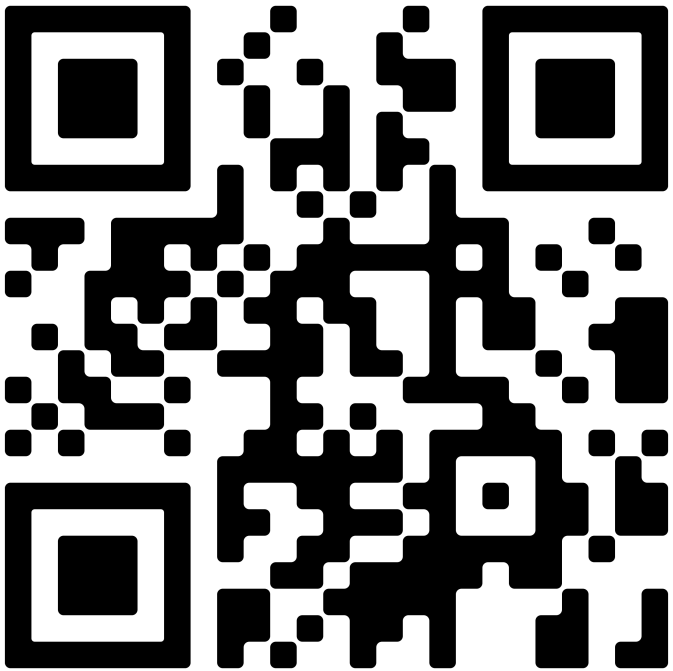

Figure: https://github.com/FJRubio67/BVSSurv

Survival Models

The typical data set

- In many areas such as medicine, biology, and engineering (reliability), scientists have access to the survival times of a group of individuals or items.
- Sample of **times to event** (possibly right-censored) $(t_1, \ldots, t_n)$ from a group of individuals.
- Vital status (or **censoring** indicators) $(\delta_1, \ldots, \delta_n)$. ($\delta_i = 1$: death, $\delta_i = 0$, right-censored/alive). Censoring may be due to random drop-out, lost to follow-up, or administrative censoring.
- In some cases, we may know some additional characteristics about the individuals, meaning we have access to **covariates** $\mathbf{x}_i = (x_{i1}, \ldots, x_ip)^\top$, (age, sex, deprivation level, comorbidities, tumour stage, ...).

Aim

- Response: Survival time (possibly censored); p covariates available $\mathbf{x}_i$.
- **Aim:** Determining what covariates have an effect (association) on the survival response.
- We need a Statistical Model.

The hazard function

- Let $T > 0$ be a (absolutely) continuous random variable with pdf $f(t)$, CDF $F(t)$, and survival function $S(t) = 1 - F(t)$.
- The **hazard function** is defined as the instantaneous risk:

$$h(t) = \lim_{dt \to 0} \frac{P[t \leq T < t + dt \mid T \geq t]}{dt} \overset{\text{homework}}{=} \frac{f(t)}{S(t)}.$$

- The function $H(t) = \int_0^t h(r)dr$ is known as the **cumulative hazard** function. $S(t) = \exp\{-H(t)\}$.
- Survival models are often formulated using the hazard function.

Proportional hazards models

- The PH model postulates that the covariates affect a "baseline hazard", by either increasing it or decreasing it. This is,

$$h_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) = h_0(t \mid \boldsymbol{\theta}) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}.$$

- The corresponding cumulative hazard function is:

$$H_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) \overset{\textit{homework}}{=} H_0(t \mid \boldsymbol{\theta}) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}.$$

The Cox PH Model and the Partial Likelihood function

- The PH model:

$$h_{PH}(t \mid \mathbf{x}_i, \boldsymbol{\beta}) = h_0(t) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\},$$

- In order to avoid misspecification of the baseline hazard (wrong model), it is often preferred to estimate it non-parametrically, while the coefficients $\boldsymbol{\beta}$ are estimated using the log partial likelihood function Cox [1972]:

$$\ell_p(\boldsymbol{\beta}) = \sum_{\delta_i=1} \mathbf{x}_i^\top \boldsymbol{\beta} - \sum_{\delta_i=1} \log\left(\sum_{k \in \mathcal{R}(t_i)} \exp\left\{\mathbf{x}_k^\top \boldsymbol{\beta}\right\}\right),$$

  where t_i, $i = 1, \ldots, n$, are the survival times, $\mathcal{R}(t) = \{i : t_i \geq t\}$ denotes the risk set at time t.

- See: https://rpubs.com/FJRubio/CPHM

Accelerated Failure Time model

- The AFT postulates that covariates affect simultaneously the time scale and the hazard scale:

$$h_{AFT}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) = h_0\left(t \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right) \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\}.$$

- The corresponding cumulative hazard function is:

$$H_{AFT}(t \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta}) \overset{\textit{homework}}{=} H_0\left(t \exp\left\{\mathbf{x}_i^\top \boldsymbol{\beta}\right\} \mid \boldsymbol{\theta}\right).$$

AFT model: interpretation

- Let $Y_i = \log(T_i)$ and $Y_i \mid \mathbf{x}_i^\top \boldsymbol{\alpha} \sim G_0(\cdot \mid \boldsymbol{\theta})$. Let g_0 be the corresponding pdf.
- The AFT model can be reformulated as a log-linear model:

$$y_i = \log(t_i) = \mathbf{x}_i^\top \boldsymbol{\alpha} + \epsilon_i, \quad i = 1, \ldots, n,$$

  where $\epsilon_i \overset{iid}{\sim} G_0(\cdot \mid \boldsymbol{\theta})$.
- Consequently, the covariates have a direct effect on the log survival time.
- See: https://github.com/FJRubio67/ShortCourseParamSurvival

Regression models: the likelihood function

- The likelihood function (for PH and AFT models) is:

$$
\begin{aligned}
L(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} f_j(t_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta})^{\delta_i} S_j(t_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta})^{1-\delta_i} \\
&\overset{\textit{homework}}{=} \prod_{i=1}^{n} h_j(t_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta})^{\delta_i} \exp\left\{-H_j(t_i \mid \mathbf{x}_i, \boldsymbol{\theta}, \boldsymbol{\beta})\right\}, \quad (*)
\end{aligned}
$$

  $j = PH, AFT$.

- This also shows that the likelihood can be characterised using the hazard function.
- Maximum Likelihood Estimators for several choices of the baseline hazard can be obtained using the R package `HazReg` (https://github.com/FJRubio67/HazReg).

Selection Methods

Stepwise (Oh no no no)

- In each step, a variable is considered for inclusion to or exclusion from the set of variables based on some specific criterion (AIC, BIC, tests, ...).
- Forward, backward, and both strategies.
- Myriad of disadvantages: lack of control on errors, *inconsistent*, lack of uncertainty quantification about the *model selection*.

Penalised methods

- LASSO + (Cox, AFT).

$$\ell_p(\boldsymbol{\beta}) - \lambda||\boldsymbol{\beta}||_1,$$
$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) - \lambda||\boldsymbol{\beta}||_1.$$

- More generally Penalty + Model.
- Advantages: tend to be fast (*e.g.* `glmnet` R package for Cox+LASSO), oracle property (under some conditions).
- Disadvantages: lack of uncertainty quantification about the *model selection*, affected when the covariates are correlated (finite sample).
- Extensions to survival knockoffs.

Spike-Slab priors

- For BVS, we need a prior, in addition to the model.
- Bayesian methods (AFT, Cox) + Spike-Slab prior.
- We introduce the variable inclusion indicators

$$\gamma_j = \begin{cases} 1 & \text{if variable } j \text{ is included} \\ 0 & \text{otherwise} \end{cases}$$

- Discrete construction (mixture prior)

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\delta_0(\beta_j) + \gamma_j N(0, \eta_j),$$

δ_0 is a mass probability at 0.

Continuous Spike-Slab priors

- Continuous construction

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j) F(0, \tilde{\eta}_j) + \gamma_j N(0, \eta_j),$$

F can be a Laplace distribution (spike-and-slab LASSO), or another scale mixture of normals.

- Variable selection can then be achieved by calculating joint posterior probabilities of vectors γ and selecting the largest one, or by calculating frequencies of inclusion of γ_j and choosing those that exceed a certain threshold.
- Advantages: tends to be fast, just need to obtain the posterior of $\gamma = (\gamma_1, \ldots, \gamma_p)$ (MCMC).
- Disadvantage: there is no posterior distribution over the model space, so model uncertainty quantification is more challenging. (Not impossible)
- See: Handbook of Bayesian Variable Selection [Tadesse and Vannucci, 2021].

Other methods

- BART: Bayesian Additive Regression Trees.
- Survival model + Shrinkage priors.
- Machine learning methods: Random survival forests, ...
- See: Handbook of Bayesian Variable Selection Tadesse and Vannucci [2021].
- General message: **No Panacea**.

Bayesian Variable Selection

Model formulation: AFT model

- The AFT model postulates

$$\log(t_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

  where ϵ_i are independent across $i = 1, \ldots, n$ with mean $E(\epsilon_i) = 0$ and variance $V(\epsilon_i) = \sigma^2$ (assumed finite).
- We will focus on the case where $\epsilon_i \sim N(0, \sigma^2)$, but other distributional assumptions are possible.

Likelihood

- It is convenient to reparameterize $\alpha = \beta/\sigma$, and $\tau = 1/\sigma$, as then the log-likelihood is concave, provided the number of uncensored individuals is greater than the number of model parameters ($n_o \geq p$) and that X_o has full column rank (the design matrix associated to the uncensored observations).
- The log-likelihood is

$$\begin{aligned} \ell(\alpha, \tau) &= -\frac{n_o}{2} \log\left(\frac{2\pi}{\tau^2}\right) - \frac{1}{2} \sum_{\delta_i=1} (\tau y_i - x_i^\top \alpha)^2 \\ &+ \sum_{\delta_i=0} \log\left\{\Phi\left(x_i^\top \alpha - \tau y_i\right)\right\}, \end{aligned} \tag{1}$$

Model Selection

- Our goal is model selection, which we formalize as choosing among two possibilities

$$\gamma_j = \begin{cases} 0, & \text{if } \beta_j = 0, \\ 1, & \text{if } \beta_j \neq 0, \end{cases}$$

  corresponding to no effect, or the inclusion of a linear effect of each covariate $j = 1, \ldots, p$. There are extensions to selection of non-linear effects but they are beyond the aims of this [Rossell and Rubio, 2023].
- That is, $\gamma = (\gamma_1, \ldots, \gamma_p)$ determines what covariates enter the model and their effect, and there are 2^p total models to consider.

Bayesian Variable Selection

The posterior model probabilities

$$\pi(\gamma \mid y) = \frac{p(y \mid \gamma)\pi(\gamma)}{\sum_{\gamma} p(y \mid \gamma)\pi(\gamma)}, \tag{2}$$

where $\pi(\gamma)$ is the model prior probability, and

$$p(y \mid \gamma) = \int p(y \mid \alpha_\gamma, \tau)\pi(\alpha_\gamma, \tau \mid \gamma) d\alpha_\gamma d\tau,$$

the integrated likelihood $p(y \mid \alpha_\gamma, \tau)$ with respect to a prior density $\pi(\alpha_\gamma, \tau \mid \gamma)$.
Not an easy task!

BVS

- One may choose the model with highest $\pi(\gamma \mid y)$, variables with high marginal posterior probabilities $\pi(\gamma_j \neq 0 \mid y)$ and,
- when the interest is in prediction, use Bayesian model averaging where models are weighted according to $\pi(\gamma \mid y)$,
- or alternatively choosing a sparse model giving similar predictions.

Either way $\pi(\gamma \mid y)$ are critical for inference, hence the importance to understand their behavior.

Priors

Two priors: local and non-local

Idea: g-priors or non-local priors for α.

$$\pi_L(\alpha_\gamma, \tau) = \prod_{\gamma_j=1} N\left(\alpha_j; 0, g_L n/(x_j^\top x_j)\right) \pi(\tau)$$

$$\pi_M(\alpha_\gamma, \tau) = \prod_{\gamma_j=1} \frac{\alpha_j^2}{g_M} N\left(\alpha_j; 0, g_M\right) \pi(\tau).$$

Priors on the precision and the model

- By default, we consider independent Beta-Binomial priors

$$\pi(\gamma) = \text{BetaBin}(p_\gamma; p, a_1, b_1),$$

  where $\text{BetaBin}(z; p, a, b)$ is Beta-Binomial distribution.
- Any model such that the number of parameters is $p_\gamma > n$ is assigned $\pi(\gamma) = 0$, as it would result in data interpolation.
- $\pi(\tau) = 2\tau^{-3}\text{IG}(\tau^{-2}; a_\tau/2, b_\tau/2)$, and IG denotes the inverse gamma density,
- and $g_L, g_M, a_\tau, b_\tau \in \mathbb{R}_+$ are given dispersion parameters: Not an automatic method!

Local prior

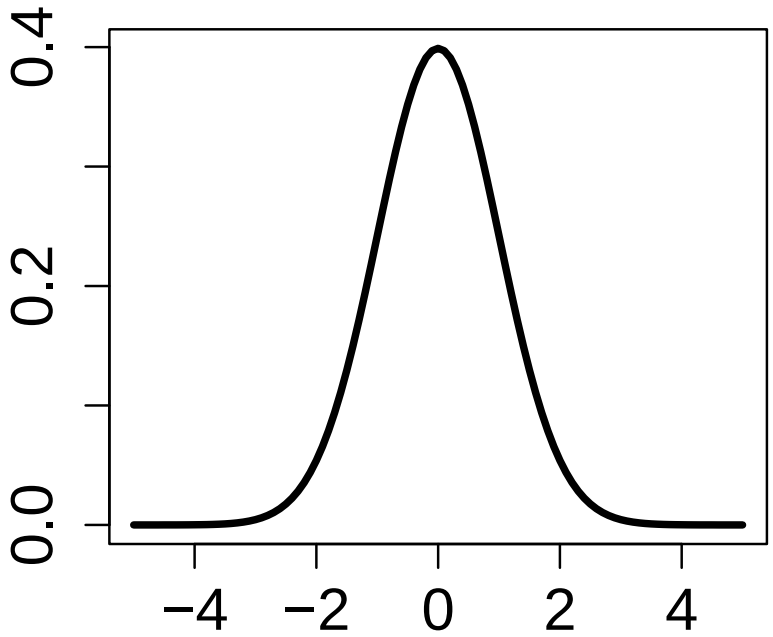

Figure: g-prior

Local prior

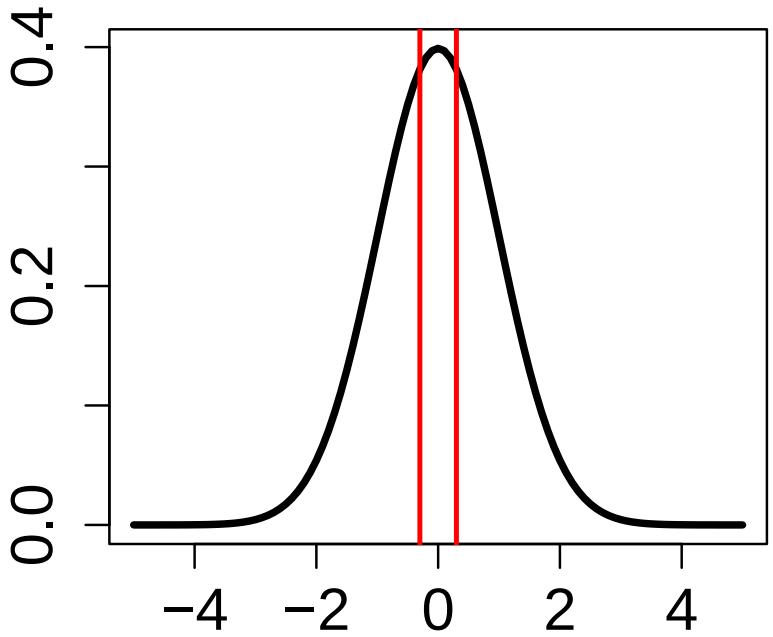

Figure: g-prior

Non-local prior

Johnson and Rossell [2010]

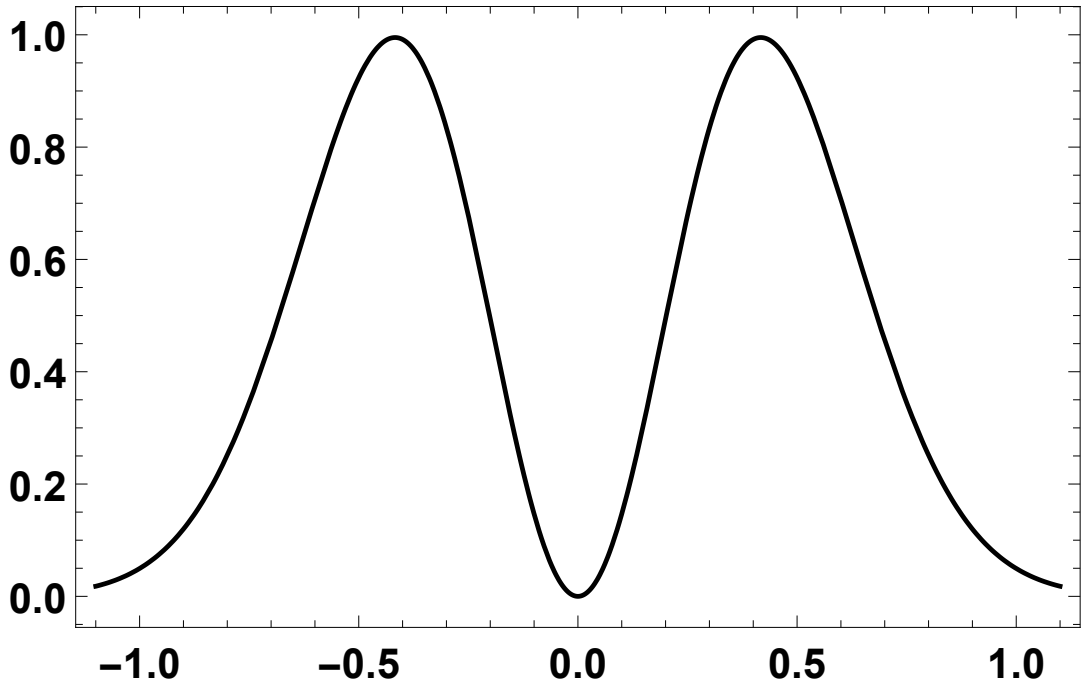

Figure: “Bayes factors that are obtained by using local alternative priors exhibit a disturbing large sample property. As the sample size n increases, they accumulate evidence much more rapidly in favour of true alternative models than in favour of true null models.”

Computations

Computations: Approximations and Model Space Exploration

- We approximate the marginal likelihood $p(y \mid \gamma)$ using a **Laplace approximation** (tricks to recycle and make more efficient some calculations).

$$\widehat{p}(y \mid \gamma) = \exp\{\ell(\tilde{\eta}_\gamma) + \log \pi(\tilde{\eta}_\gamma)\}(2\pi)^{d_\gamma/2} \left| H(\tilde{\eta}_\gamma) + \nabla^2 \log \pi(\tilde{\eta}_\gamma) \right|^{-1/2},$$

  where $\tilde{\eta}_\gamma = \arg\max_{\eta_\gamma} \ell(\eta_\gamma) + \log \pi(\eta_\gamma)$ is the maximum a posteriori under prior $\pi(\eta_\gamma)$. **Coordinate Descent Algorithm** (among others).
- Not covered: Approximate Laplace Approximations [Rossell et al., 2021].

Approximations and Model Space Exploration

- **Model exploration**: Gibbs sampling (MCMC in general). Active research area. For instance, a simple algorithm:
  1. Initialise $\gamma = \gamma_0$.
  2. Update $\gamma_j^{(t)} = 1$ with probability

$$\frac{p(\gamma_1^{(t)}, \ldots, \gamma_{j-1}^{(t)}, \gamma_j^{(t-1)}, \gamma_{j+1}^{(t)}, \ldots, \gamma_p^{(t)} \mid \mathbf{y})}{\sum_{\gamma_j=0}^{1} p(\gamma_1^{(t)}, \ldots, \gamma_{j-1}^{(t)}, \gamma_j, \gamma_{j+1}^{(t)}, \ldots, \gamma_p^{(t)} \mid \mathbf{y})}$$

Theory

To interpret the results in the M-Open scenario, we need to define the expected log-likelihood under the data-generating F_0.

$$M(\eta_\gamma) = \frac{1}{n}\mathbb{E}_{F_0}(\ell(\eta_\gamma))$$

Under minimal conditions, $M(\eta_\gamma)$ has a unique maximiser, denoted by $\eta_\gamma^* = (\alpha_\gamma^*, \tau_\gamma^*)$. M is affected by both, **the survival process and the censoring mechanism**.

Theory: take-home messages

- $\eta_\gamma^* = \text{argmax}_{\Gamma_\gamma} M(\eta_\gamma)$ is unique and $\widehat{\eta}_\gamma \xrightarrow{P} \eta_\gamma^*$ as $n \to \infty$.
- The Laplace approximation has a relative error converging to zero as $n \to \infty$.
- $\pi(\gamma^* \mid y) \xrightarrow{P} 1$. Asymptotically selects (both priors) the model γ^* of smallest dimension maximising $M(\eta_\gamma)$.
- This implies that the highest posterior probability model consistently selects γ^*, and that including covariates with marginal posterior probability $\pi(\gamma_j^* \mid y) > t$, for any fixed threshold t, also leads to consistent selection.
- Bayesian model selection in the AFT model asymptotically returns the smallest γ^* that minimises the KL divergence between the true model and the AFT model.

Examples

GitHub repository

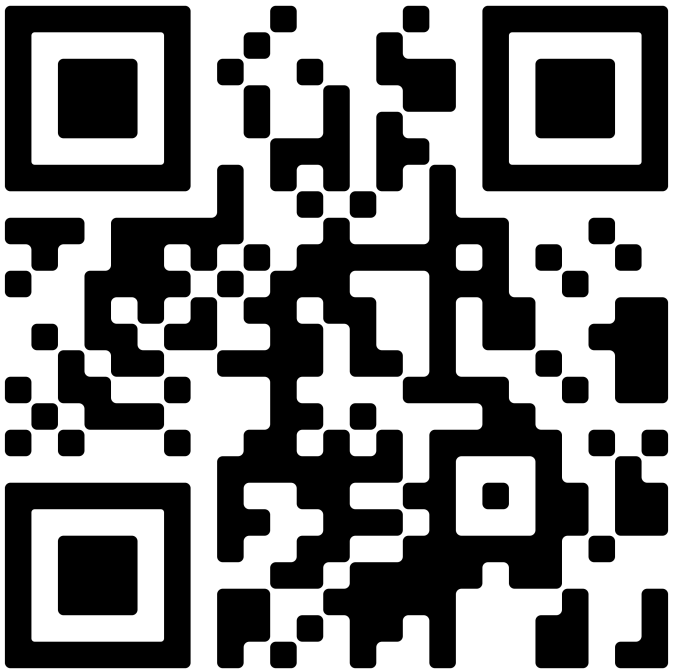

Figure: https://github.com/FJRubio67/BVSSurv

Example 1: Simulated data

- `Simulated data from an accelerated failure time model`
- $n = 250$, $p = 10$, 47% censoring.
- R code.

  https://rpubs.com/FJRubio/BVSSurvExample1

- Sensitivity analyses: Nikooienejad et al. [2020], Simon et al. [2011], Yi et al. [2019]

Example 2: flchain data set

- `This is a stratified random sample containing 1/2 of the subjects from a study of the relationship between serum free light chain (FLC) and mortality.`
- $n = 6524$, $p = 5$, 70% censoring.
- R code.

  https://rpubs.com/FJRubio/BVSSurvExample2

- Homework: answer the open question explained in the help file.

Example 3: colon cancer data set

- `Association of 172 genes + TFG-B (growth factor) + tumour stage with colon cancer survival`
- $n = 260$, $p = 175$, 80% censoring.
- R code.

  https://rpubs.com/FJRubio/BVSSurvExample3

- Homework: Stratified analysis.

Example 4: nki70 data set

- `144 lymph node positive breast cancer patients on metastasis-free survival, 5 clinical risk factors, and gene expression measurements of 70 genes found to be prognostic for metastasis-free survival in an earlier study.`
- $n = 144$, $p = 75$, 66% censoring.
- R code.

  https://rpubs.com/FJRubio/BVSSurvExample4

- Small sample, moderate dimension, high censoring.
- Homework: Reflect on reasons for differences (prior calibration? Model misspecification?, All?).
- Homework 2: correct data preparation (using factors and dummy variables instead of numeric).

Take-home messages

- We have reviewed several (Bayesian) variable selection methods for survival data.
- We have studied methodology for the selection of variables in the context of AFT models. Implementation in the R package `mombf`.
- Model misspecification has a finite-sample and asymptotic effects on the performance of variable selection.
- Censoring also has an effect in the finite-sample scenario: should we increase the sample size or the follow-up to improve power of BVS?
- Bayesian variable selection is not automatic: prior calibration is crucial in the finite sample scenario.

Additional software tools

- BVS for Cox model and non-local prior: `BVSNLP` R package.
- Spike and Slab LASSO (only posterior modes): `BhGLM` R package.
- BART: `BART` R package.
- Random Survival Forests: `randomForestSRC` R package.
- Cox-LASSO: `glmnet` R package.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34 (2):187–220, 1972.

V.E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B*, 72(2):143–170, 2010.

A. Nikooienejad, W. Wang, and V.E. Johnson. Bayesian variable selection for survival data using inverse moment priors. *Annals of Applied Statistics*, 14:809–828, 2020.

D. Rossell and F.J. Rubio. Additive Bayesian variable selection under censoring and misspecification. *Statistical Science*, 38:13–29, 2023.

D. Rossell, O. Abril, and A. Bhattacharya. Approximate Laplace approximations for scalable model selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83 (4):853–879, 2021.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1, 2011.

M.G. Tadesse and M. Vannucci. *Handbook of Bayesian variable selection*. CRC Press, 2021.

N. Yi, Z. Tang, X. Zhang, and B. Guo. Bhglm: Bayesian hierarchical glms and survival models, with applications to genomics and epidemiology. *Bioinformatics*, 35(8):1419–1421, 2019.