A model selection approach for variable selection with censored data

Maria Eugenia Castellanos[1], Gonzalo Garcia-Donato[2] and Stefano Cabras[3]

[1] U. Rey Juan Carlos (Spain), [2] U. de Castilla-La Mancha (Spain), [3] U. Carlos III de Madrid (Spain)

June 2019 - OBayes19 Conference

Highlights on this talk

- We discuss about the convenience of correcting model selection priors in the presence of censored observations,

Highlights on this talk

- We discuss about the convenience of correcting model selection priors in the presence of censored observations,
- We derive extensions of g-priors in this scenario, that include a new definition of sample size,

Highlights on this talk

- We discuss about the convenience of correcting model selection priors in the presence of censored observations,
- We derive extensions of g-priors in this scenario, that include a new definition of sample size,
- We evaluate our proposal under the scrutiny of Predictive Matching arguments,

Highlights on this talk

- We discuss about the convenience of correcting model selection priors in the presence of censored observations,
- We derive extensions of g-priors in this scenario, that include a new definition of sample size,
- We evaluate our proposal under the scrutiny of Predictive Matching arguments,
- Illustrate our methodology using a real data set of a breast cancer registry.

Highlights on this talk

- We discuss about the convenience of correcting model selection priors in the presence of censored observations,
- We derive extensions of g-priors in this scenario, that include a new definition of sample size,
- We evaluate our proposal under the scrutiny of Predictive Matching arguments,
- Illustrate our methodology using a real data set of a breast cancer registry.

And we are very Bayesian and very objective, and do not allow ourselves using any type of sample information to define our priors.

The model selection approach

In variable selection, we have an initial set of potential explanatory variables:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$$

and we have to select those that are relevant to explain the variability of a response variable Y.

The model selection approach

In variable selection, we have an initial set of potential explanatory variables:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$$

and we have to select those that are relevant to explain the variability of a response variable Y.

• Within the model selection approach, the answer to variable selection is obtained from the posterior probabilities of the 2^k possible models:

$$p(\mathcal{M}_\gamma \mid \boldsymbol{y}),\ \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k),\ \gamma_i \in \{0, 1\}.$$

The model selection approach

In variable selection, we have an initial set of potential explanatory variables:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$$

and we have to select those that are relevant to explain the variability of a response variable Y.

• Within the model selection approach, the answer to variable selection is obtained from the posterior probabilities of the 2^k possible models:

$$p(\mathcal{M}_\gamma \mid \boldsymbol{y}),\ \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k),\ \gamma_i \in \{0, 1\}.$$

• This talk concerns the assignment of prior distributions for the specific parameters within each model $\mathcal{M}_\gamma$. It suffices to present the problem as if only two models (the full and the null) were entertained. The proposal automatically generalizes to the 2^k models situation.

Model selection within the linear model: basic formula

In the regular linear model, the model that contains all k covariates (full model) is:

$$\mathcal{M}(\boldsymbol{y} \mid \boldsymbol{\beta}, \beta_0, \sigma) : y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \;\; \epsilon_i \sim N(0, 1),$$

where $\widetilde{\boldsymbol{x}}_i \in \mathcal{R}^k$ is the vector of centered (values) of covariates for sample i.

Model selection within the linear model: basic formula

In the regular linear model, the model that contains all k covariates (full model) is:

$$\mathcal{M}(\boldsymbol{y} \mid \boldsymbol{\beta}, \beta_0, \sigma) : y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \ \epsilon_i \sim N(0, 1),$$

where $\widetilde{\boldsymbol{x}}_i \in \mathcal{R}^k$ is the vector of centered (values) of covariates for sample i.

The null model:

$$\mathcal{M}_0(\boldsymbol{y} \mid \beta_0, \sigma) : y_i = \beta_0 + \sigma\epsilon_i.$$

Model selection within the linear model: basic formula

In the regular linear model, the model that contains all k covariates (full model) is:

$$\mathcal{M}(\boldsymbol{y} \mid \boldsymbol{\beta}, \beta_0, \sigma) : y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \;\; \epsilon_i \sim N(0,1),$$

where $\widetilde{\boldsymbol{x}}_i \in \mathcal{R}^k$ is the vector of centered (values) of covariates for sample i.

The null model:

$$\mathcal{M}_0(\boldsymbol{y} \mid \beta_0, \sigma) : y_i = \beta_0 + \sigma\epsilon_i.$$

Posterior probabilities are

$$p(\mathcal{M} \mid \boldsymbol{y}) = \frac{m(\boldsymbol{y})p(\mathcal{M})}{m(\boldsymbol{y})p(\mathcal{M}) + m_0(\boldsymbol{y})p(\mathcal{M}_0)}, \;\; p(\mathcal{M}_0 \mid \boldsymbol{y}) = \frac{m_0(\boldsymbol{y})p(\mathcal{M}_0)}{m(\boldsymbol{y})p(\mathcal{M}) + m_0(\boldsymbol{y})p(\mathcal{M}_0)}.$$

Where:

$$m(\boldsymbol{y}) = \int \mathcal{M}(\boldsymbol{y} \mid \boldsymbol{\beta}, \beta_0, \sigma)\pi(\boldsymbol{\beta}, \beta_0, \sigma)\, d\boldsymbol{\beta}\, d\beta_0 d\sigma, \;\; m_0(\boldsymbol{y}) = \int \mathcal{M}_0(\boldsymbol{y} \mid \beta_0, \sigma)\pi_0(\beta_0, \sigma)\, d\beta_0 d\sigma$$

The ratio $B = m(\boldsymbol{y})/m_0(\boldsymbol{y})$ is the Bayes factor (to the null).

For $\{p(\mathcal{M}), p(\mathcal{M}_0)\}$ we use $p(\mathcal{M}) = p(\mathcal{M}_0) = 0.5$ and, in the case of variable selection, the prior studied in Scott and Berger (2010). The focus on this talk is on the priors for parameters within each model:

$$\pi_0(\beta_0, \sigma) \;\; \pi(\boldsymbol{\beta}, \beta_0, \sigma).$$

Variable selection priors have a common starting point

$$\pi_0(\beta_0, \sigma), \text{ and } \pi(\boldsymbol{\beta}, \beta_0, \sigma) = \pi_0(\beta_0, \sigma) \times \pi(\boldsymbol{\beta} \mid \beta_0, \sigma),$$

where

- $\pi_0(\beta_0, \sigma)$ is an objective estimation prior (normally $\pi_0(\beta_0, \sigma) = \sigma^{-1}$ or vague versions of it) and
- the conditional prior for the specific parameters: $\pi(\boldsymbol{\beta} \mid \beta_0, \sigma)$ is a proper prior.

Variable selection priors have a common starting point

$$\pi_0(\beta_0, \sigma), \text{ and } \pi(\boldsymbol{\beta}, \beta_0, \sigma) = \pi_0(\beta_0, \sigma) \times \pi(\boldsymbol{\beta} \mid \beta_0, \sigma),$$

where

- $\pi_0(\beta_0, \sigma)$ is an objective estimation prior (normally $\pi_0(\beta_0, \sigma) = \sigma^{-1}$ or vague versions of it) and
- the conditional prior for the specific parameters: $\pi(\boldsymbol{\beta} \mid \beta_0, \sigma)$ is a proper prior.

One of the most popular approaches to specify $\pi(\boldsymbol{\beta} \mid \beta_0, \sigma)$ dates back to Zellner and Siow (1980) based on the previous work by Jeffreys (1961). It has been extended in various ways with important contributions in the literature: Zellner (1986); Fernández et al. (2001); Liang et al. (2008); Bayarri et al. (2012) (just to mention some).

Variable selection priors have a common starting point

$$\pi_0(\beta_0, \sigma), \text{ and } \pi(\boldsymbol{\beta}, \beta_0, \sigma) = \pi_0(\beta_0, \sigma) \times \pi(\boldsymbol{\beta} \mid \beta_0, \sigma),$$

where

- $\pi_0(\beta_0, \sigma)$ is an objective estimation prior (normally $\pi_0(\beta_0, \sigma) = \sigma^{-1}$ or vague versions of it) and
- the conditional prior for the specific parameters: $\pi(\boldsymbol{\beta} \mid \beta_0, \sigma)$ is a proper prior.

One of the most popular approaches to specify $\pi(\boldsymbol{\beta} \mid \beta_0, \sigma)$ dates back to Zellner and Siow (1980) based on the previous work by Jeffreys (1961). It has been extended in various ways with important contributions in the literature: Zellner (1986); Fernández et al. (2001); Liang et al. (2008); Bayarri et al. (2012) (just to mention some).

This myriad of proposals have been named g-priors or conventional priors (Berger and Pericchi, 2001; Bayarri and García-Donato, 2007) and have in common special features that now I summarize.

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}), \;\; g \sim \pi(g).$$

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}), \;\; g \sim \pi(g).$$

Here $\pi(g)$ is either degenerate to a fixed value (e.g. $g = 1$ in BIC) or provides the prior with convenient flat tails (e.g. Inverse Gamma) giving rise to the g-priors spectrum.

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}),\ \ g \sim \pi(g).$$

Here $\pi(g)$ is either degenerate to a fixed value (e.g. $g = 1$ in BIC) or provides the prior with convenient flat tails (e.g. Inverse Gamma) giving rise to the g-priors spectrum.
By default, we use the Robust prior for $\pi(g)$ (but any other can be easily implemented).

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}), \; g \sim \pi(g).$$

Here $\pi(g)$ is either degenerate to a fixed value (e.g. $g = 1$ in BIC) or provides the prior with convenient flat tails (e.g. Inverse Gamma) giving rise to the g-priors spectrum. By default, we use the Robust prior for $\pi(g)$ (but any other can be easily implemented).

With respect to $\boldsymbol{\Sigma}$, g priors propose a quite particular form:

$$\boldsymbol{\Sigma} = n\sigma^2(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}, \quad \widetilde{\boldsymbol{X}}^\top = (\widetilde{\boldsymbol{x}}_1^\top \cdots \widetilde{\boldsymbol{x}}_n^\top) \text{ (the centered design matrix).}$$

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}),\ \ g \sim \pi(g).$$

Here $\pi(g)$ is either degenerate to a fixed value (e.g. $g = 1$ in BIC) or provides the prior with convenient flat tails (e.g. Inverse Gamma) giving rise to the g-priors spectrum. By default, we use the Robust prior for $\pi(g)$ (but any other can be easily implemented).

With respect to $\boldsymbol{\Sigma}$, g priors propose a quite particular form:

$$\boldsymbol{\Sigma} = n\sigma^2(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1},\ \ \widetilde{\boldsymbol{X}}^\top = (\widetilde{\boldsymbol{x}}_1^\top \cdots \widetilde{\boldsymbol{x}}_n^\top)\ \text{(the centered design matrix).}$$

This debatable choice (versus eg independent prior):

- Makes the $\boldsymbol{\beta}$'s depends on the $X's$, varying inversely proportional to $Var(X)$'s,

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}), \;\; g \sim \pi(g).$$

Here $\pi(g)$ is either degenerate to a fixed value (e.g. $g = 1$ in BIC) or provides the prior with convenient flat tails (e.g. Inverse Gamma) giving rise to the g-priors spectrum. By default, we use the Robust prior for $\pi(g)$ (but any other can be easily implemented).

With respect to $\boldsymbol{\Sigma}$, g priors propose a quite particular form:

$$\boldsymbol{\Sigma} = n\sigma^2(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}, \;\; \widetilde{\boldsymbol{X}}^\top = (\widetilde{\boldsymbol{x}}_1^\top \cdots \widetilde{\boldsymbol{x}}_n^\top) \text{ (the centered design matrix).}$$

This debatable choice (versus eg independent prior):

- Makes the β's depends on the $X's$, varying inversely proportional to $Var(X)$'s,
- is inspired by the expected Fisher information matrix,

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}), \; g \sim \pi(g).$$

Here $\pi(g)$ is either degenerate to a fixed value (e.g. $g = 1$ in BIC) or provides the prior with convenient flat tails (e.g. Inverse Gamma) giving rise to the g-priors spectrum. By default, we use the Robust prior for $\pi(g)$ (but any other can be easily implemented).

With respect to $\boldsymbol{\Sigma}$, g priors propose a quite particular form:

$$\boldsymbol{\Sigma} = n\sigma^2(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}, \quad \widetilde{\boldsymbol{X}}^\top = (\widetilde{\boldsymbol{x}}_1^\top \cdots \widetilde{\boldsymbol{x}}_n^\top) \text{ (the centered design matrix).}$$

This debatable choice (versus eg independent prior):

- Makes the $\boldsymbol{\beta}$'s depends on the $X's$, varying inversely proportional to $Var(X)$'s,
- is inspired by the expected Fisher information matrix,
- contains the factor n that makes it of unitary size.

Radiography of conventional g-priors

- Within g priors:

$$\pi(\boldsymbol{\beta} \mid \beta_0, \sigma, g) = N(0, g\boldsymbol{\Sigma}), \;\; g \sim \pi(g).$$

Here $\pi(g)$ is either degenerate to a fixed value (e.g. $g = 1$ in BIC) or provides the prior with convenient flat tails (e.g. Inverse Gamma) giving rise to the g-priors spectrum. By default, we use the Robust prior for $\pi(g)$ (but any other can be easily implemented).

With respect to $\boldsymbol{\Sigma}$, g priors propose a quite particular form:

$$\boldsymbol{\Sigma} = n\sigma^2(\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}, \;\; \widetilde{\boldsymbol{X}}^\top = (\widetilde{\boldsymbol{x}}_1^\top \cdots \widetilde{\boldsymbol{x}}_n^\top) \text{ (the centered design matrix).}$$

This debatable choice (versus eg independent prior):

- Makes the $\boldsymbol{\beta}$'s depends on the $X's$, varying inversely proportional to $Var(X)$'s,
- is inspired by the expected Fisher information matrix,
- contains the factor n that makes it of unitary size.
- Particular properties: exact null predictive matching (Bayarri et al., 2012); Group invariant (Consonni et al, 2019).

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i,\ \ i = 1, 2, \ldots, n$$

but

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \;\; i = 1, 2, \ldots, n$$

but

- we observe y_i only if $y_i < c_i$, in which case we record $\delta_i = 1$,

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \ \ i = 1, 2, \ldots, n$$

but

- we observe y_i only if $y_i < c_i$, in which case we record $\delta_i = 1$,
- if $y_i \geq c_i$, we record $\delta_i = 0$.

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma \epsilon_i, \;\; i = 1, 2, \ldots, n$$

but

- we observe y_i only if $y_i < c_i$, in which case we record $\delta_i = 1$,
- if $y_i \geq c_i$, we record $\delta_i = 0$.

The c_i are censoring times and we assume that for all people in the study

$$c_1, c_2, \ldots, c_n$$

are known.

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \;\; i = 1, 2, \ldots, n$$

but

- we observe y_i only if $y_i < c_i$, in which case we record $\delta_i = 1$,
- if $y_i \geq c_i$, we record $\delta_i = 0$.

The c_i are censoring times and we assume that for all people in the study

$$c_1, c_2, \ldots, c_n$$

are known.

Example at the end: prognosis factors for survival to breast cancer

- Observational units are women diagnosed with the disease in 2004-2013.

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma \epsilon_i, \;\; i = 1, 2, \ldots, n$$

but

- we observe y_i only if $y_i < c_i$, in which case we record $\delta_i = 1$,
- if $y_i \geq c_i$, we record $\delta_i = 0$.

The c_i are censoring times and we assume that for all people in the study

$$c_1, c_2, \ldots, c_n$$

are known.

Example at the end: prognosis factors for survival to breast cancer

- Observational units are women diagnosed with the disease in 2004-2013.
- We are interested on y_i time to death (since diagnosis), in log scale

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \;\; i = 1, 2, \ldots, n$$

but

- we observe y_i only if $y_i < c_i$, in which case we record $\delta_i = 1$,
- if $y_i \geq c_i$, we record $\delta_i = 0$.

The c_i are censoring times and we assume that for all people in the study

$$c_1, c_2, \ldots, c_n$$

are known.

Example at the end: prognosis factors for survival to breast cancer

- Observational units are women diagnosed with the disease in 2004-2013.
- We are interested on y_i time to death (since diagnosis), in log scale
- The study was planned to end on 12/31/2015. For every woman in the study

$$c_i = log\big(12/31/2015 - \text{diagnosis date}\big)$$

The linear model with censoring

Now, the model that contains the covariates is:

$$y_i = \beta_0 + \boldsymbol{\beta}^\top \widetilde{\boldsymbol{x}}_i + \sigma\epsilon_i, \;\; i = 1, 2, \ldots, n$$

but

- we observe y_i only if $y_i < c_i$, in which case we record $\delta_i = 1$,
- if $y_i \geq c_i$, we record $\delta_i = 0$.

The c_i are censoring times and we assume that for all people in the study

$$c_1, c_2, \ldots, c_n$$

are known.

Example at the end: prognosis factors for survival to breast cancer

- Observational units are women diagnosed with the disease in 2004-2013.
- We are interested on y_i time to death (since diagnosis), in log scale
- The study was planned to end on 12/31/2015. For every woman in the study

$$c_i = log\left(12/31/2015 - \text{diagnosis date}\right)$$

- Women that died before 12/31/2015 are uncensored and y_i is recorded. For the rest we only know that $y_i > c_i$.

Differing information content

A priori, the "information" content in the units $i \in \{1, 2, \ldots, n\}$ varies:

Differing information content

A priori, the "information" content in the units $i \in \{1, 2, \ldots, n\}$ varies:
Roughly speaking:

- if c_i is small then (a priori) y_i has less chances to be observed and unit i will contribute partially to the likelihood. As c_i decreases, contribution of unit i to the likelihood will be negligible (no impact on inferences).

Differing information content

A priori, the "information" content in the units $i \in \{1, 2, \ldots, n\}$ varies:
Roughly speaking:

- if c_i is small then (a priori) y_i has less chances to be observed and unit i will contribute partially to the likelihood. As c_i decreases, contribution of unit i to the likelihood will be negligible (no impact on inferences).
- As c_i gets large, unit i is expected to provide full information.

Once the experiment is finished, the data we have are:

$$(\boldsymbol{y}, \boldsymbol{\delta}) = ((y_1, \ldots, y_{n_u}), (\delta_1, \ldots, \delta_n)), \quad n_u = \#\text{uncensored observations.}$$

The rest $n_c = n - n_u$ units are censored.

Differing information content

A priori, the "information" content in the units $i \in \{1, 2, \ldots, n\}$ varies:
Roughly speaking:

- if c_i is small then (a priori) y_i has less chances to be observed and unit i will contribute partially to the likelihood. As c_i decreases, contribution of unit i to the likelihood will be negligible (no impact on inferences).
- As c_i gets large, unit i is expected to provide full information.

Once the experiment is finished, the data we have are:

$$(\boldsymbol{y}, \boldsymbol{\delta}) = ((y_1, \ldots, y_{n_u}), (\delta_1, \ldots, \delta_n)), \quad n_u = \#\text{uncensored observations.}$$

The rest $n_c = n - n_u$ units are censored.

A compact expression of the model is:

$$\mathcal{M}(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \times Pr\big(N_{n_c}(\mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_c\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) > \boldsymbol{c}_c\big),$$

where

Differing information content

A priori, the "information" content in the units $i \in \{1, 2, \ldots, n\}$ varies:
Roughly speaking:

- if c_i is small then (a priori) y_i has less chances to be observed and unit i will contribute partially to the likelihood. As c_i decreases, contribution of unit i to the likelihood will be negligible (no impact on inferences).
- As c_i gets large, unit i is expected to provide full information.

Once the experiment is finished, the data we have are:

$$(\boldsymbol{y}, \boldsymbol{\delta}) = ((y_1, \ldots, y_{n_u}), (\delta_1, \ldots, \delta_n)), \quad n_u = \#\text{uncensored observations.}$$

The rest $n_c = n - n_u$ units are censored.

A compact expression of the model is:

$$\mathcal{M}(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \times Pr\big(N_{n_c}(\mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_c\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) > \boldsymbol{c}_c\big),$$

where

$$\widetilde{\boldsymbol{X}}^{\mathsf{T}} = (\widetilde{\boldsymbol{X}}_u^{\mathsf{T}}, \widetilde{\boldsymbol{X}}_c^{\mathsf{T}}), \quad \boldsymbol{c}^{\mathsf{T}} = (\boldsymbol{c}_u^{\mathsf{T}}, \boldsymbol{c}_c^{\mathsf{T}})$$

Objective model selection priors (unlike estimation priors) are partially proper in a way that depends on the observed values of covariates $\boldsymbol{x}_i$ (and on n):

Objective model selection priors (unlike estimation priors) are partially proper in a way that depends on the observed values of covariates $\boldsymbol{x}_i$ (and on n):

- Explicitly like with the g-priors,

Objective model selection priors (unlike estimation priors) are partially proper in a way that depends on the observed values of covariates $\boldsymbol{x}_i$ (and on n):

- Explicitly like with the g-priors,
- More subtly through a pre-processing (eg. standardization).

Objective model selection priors (unlike estimation priors) are partially proper in a way that depends on the observed values of covariates $\boldsymbol{x}_i$ (and on n):

- Explicitly like with the g-priors,
- More subtly through a pre-processing (eg. standardization).

Main questions

In the presence of censoring,

- should we modify the way $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ contribute to the prior?

Objective model selection priors (unlike estimation priors) are partially proper in a way that depends on the observed values of covariates $\boldsymbol{x}_i$ (and on n):

- Explicitly like with the g-priors,
- More subtly through a pre-processing (eg. standardization).

Main questions

In the presence of censoring,

- should we modify the way $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ contribute to the prior?
- n is n?

Objective model selection priors (unlike estimation priors) are partially proper in a way that depends on the observed values of covariates $\boldsymbol{x}_i$ (and on n):

- Explicitly like with the g-priors,
- More subtly through a pre-processing (eg. standardization).

Main questions

In the presence of censoring,

- should we modify the way $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ contribute to the prior?
- n is n?

In general, objective priors are directly imported from the uncensored literature (Sha et al. (2006) with spike-and-slab priors or Nikooienejad et al. (2018) with non-local priors).

Interestingly, other authors have argued about the need to rethink the notion of sample size to define their priors:

- Volinsky and Raftery (2000) propose using a version of BIC that uses the number of uncensored observations n_u (instead of n).

Objective model selection priors (unlike estimation priors) are partially proper in a way that depends on the observed values of covariates $\boldsymbol{x}_i$ (and on n):

- Explicitly like with the g-priors,
- More subtly through a pre-processing (eg. standardization).

Main questions

In the presence of censoring,

- should we modify the way $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ contribute to the prior?
- n is n?

In general, objective priors are directly imported from the uncensored literature (Sha et al. (2006) with spike-and-slab priors or Nikooienejad et al. (2018) with non-local priors).

Interestingly, other authors have argued about the need to rethink the notion of sample size to define their priors:

- Volinsky and Raftery (2000) propose using a version of BIC that uses the number of uncensored observations n_u (instead of n).
- Similarly, Held et al. (2016), make an implicit use of g-priors (with test-based Bayes factors) discussing on the convenience of using n_u to scale the prior covariance matrix.

As in the conventional approach without censoring we use:

$$\pi_0(\beta_0, \sigma) = \sigma^{-1},\ \pi(\boldsymbol{\beta}, \beta_0, \sigma) = \sigma^{-1} \times \int N(\boldsymbol{\beta} \mid \mathbf{0}, g\boldsymbol{\Sigma})\pi(g)dg,$$

but which covariance $\boldsymbol{\Sigma}$?

As in the conventional approach without censoring we use:

$$\pi_0(\beta_0, \sigma) = \sigma^{-1},\ \pi(\boldsymbol{\beta}, \beta_0, \sigma) = \sigma^{-1} \times \int N(\boldsymbol{\beta} \mid \mathbf{0}, g\boldsymbol{\Sigma})\pi(g)dg,$$

but which covariance $\boldsymbol{\Sigma}$?

• The default choice is to use **All** units equally: $\boldsymbol{\Sigma}^{All} = n\sigma^2(\widetilde{\boldsymbol{X}}^{\top}\widetilde{\boldsymbol{X}})^{-1}$, but this may have unexpected consequences that we illustrate in an extreme situation

Study (dramatization of a possible situation)

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).
- People that haven't had any symptoms at the end of the study are censored.

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).
- People that haven't had any symptoms at the end of the study are censored.
- We investigate the relation of Y with long-term memory capacity (say x_i measures the quantity of details from childhood one is able to remember).

Study (dramatization of a possible situation)

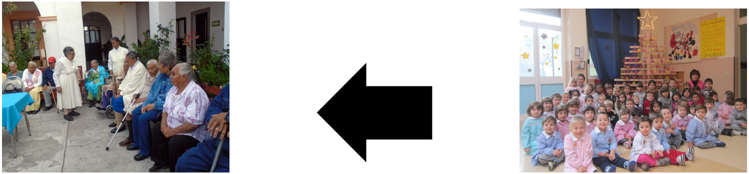
- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).
- People that haven't had any symptoms at the end of the study are censored.
- We investigate the relation of Y with long-term memory capacity (say x_i measures the quantity of details from childhood one is able to remember).

We plan a visit to an old people's home to enroll persons for the study

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).
- People that haven't had any symptoms at the end of the study are censored.
- We investigate the relation of Y with long-term memory capacity (say x_i measures the quantity of details from childhood one is able to remember).

We plan a visit to an old people's home to enroll persons for the study

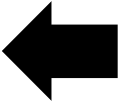


the same day that there was a school visit to the center.

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).
- People that haven't had any symptoms at the end of the study are censored.
- We investigate the relation of Y with long-term memory capacity (say x_i measures the quantity of details from childhood one is able to remember).

We plan a visit to an old people's home to enroll persons for the study



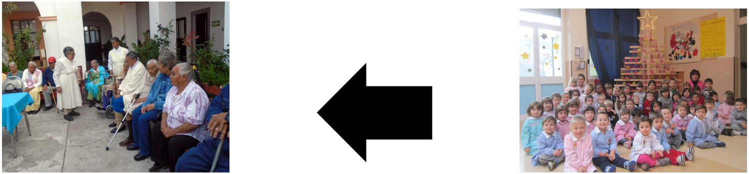
the same day that there was a school visit to the center.

• The person collecting data is paid per unit enrolled and take data from old people as well as young kids.

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).
- People that haven't had any symptoms at the end of the study are censored.
- We investigate the relation of Y with long-term memory capacity (say x_i measures the quantity of details from childhood one is able to remember).

We plan a visit to an old people's home to enroll persons for the study



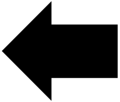
the same day that there was a school visit to the center.

• The person collecting data is paid per unit enrolled and take data from old people as well as young kids. 😱

Study (dramatization of a possible situation)

- Variable of interest (Y_i): age (in years) of appearance of early symptoms of dementia,
- Censoring time c_i is age at the end of study (known for all units).
- People that haven't had any symptoms at the end of the study are censored.
- We investigate the relation of Y with long-term memory capacity (say x_i measures the quantity of details from childhood one is able to remember).

We plan a visit to an old people's home to enroll persons for the study



the same day that there was a school visit to the center.

• The person collecting data is paid per unit enrolled and take data from old people as well as young kids.

• After 10 years the study ends. Some of the old people have shown symptoms, but all kids are censored (with very small censoring times c_i).

About the default choice: $\boldsymbol{\Sigma}^{all}$

- The likelihood

$$\mathcal{M}(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \times Pr\big(N_{n_c}(\mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_c\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) > \boldsymbol{c}_c\big),$$

About the default choice: $\boldsymbol{\Sigma}^{all}$

- The likelihood

$$\mathcal{M}(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \times 1$$

is essentially based on the old people.

About the default choice: $\boldsymbol{\Sigma}^{all}$

- The likelihood

$$\mathcal{M}(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u \boldsymbol{\beta}, \sigma^2 \boldsymbol{I}) \times 1$$

  is essentially based on the old people.

- Then, in agreement with g priors, we should use $\boldsymbol{\Sigma}^{uncens}$, a matrix based on $\widetilde{\boldsymbol{X}}_u$, the covariates for the old people.

About the default choice: $\boldsymbol{\Sigma}^{all}$

- The likelihood

$$\mathcal{M}(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \times 1$$

  is essentially based on the old people.

- Then, in agreement with g priors, we should use $\boldsymbol{\Sigma}^{uncens}$, a matrix based on $\widetilde{\boldsymbol{X}}_u$, the covariates for the old people.
- Nevertheless, the default choice (using all units) is $\boldsymbol{\Sigma}^{All}$.

About the default choice: $\boldsymbol{\Sigma}^{all}$

- The likelihood

$$\mathcal{M}(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \boldsymbol{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) \times 1$$

  is essentially based on the old people.

- Then, in agreement with g priors, we should use $\boldsymbol{\Sigma}^{uncens}$, a matrix based on $\widetilde{\boldsymbol{X}}_u$, the covariates for the old people.
- Nevertheless, the default choice (using all units) is $\boldsymbol{\Sigma}^{All}$. Clearly

$$Var(X_{all}) >> Var(X_{uncens}) \rightarrow \boldsymbol{\Sigma}^{All} << \boldsymbol{\Sigma}^{uncens},$$

  implying a more precise prior than it should, leading to conservative Bayes factors.

Observations

- The previous example is an extreme situation, but it is not unusual at all that $Var(X_{all}) >> Var(X_{uncens})$ (eg. the famous CHS and PBC survival datasets in Volinsky and Raftery, 2000).

Observations

- The previous example is an extreme situation, but it is not unusual at all that $Var(X_{all}) >> Var(X_{uncens})$ (eg. the famous CHS and PBC survival datasets in Volinsky and Raftery, 2000).
- A default usage of conventional prior is not adaptive and may have a potential impact on the Bayes factors. Most of the model selection priors in the literature would have a similar misbehavior as they depend (implicitly) on the values of covariates (eg. through standardization).

Observations

- The previous example is an extreme situation, but it is not unusual at all that $Var(X_{all}) >> Var(X_{uncens})$ (eg. the famous CHS and PBC survival datasets in Volinsky and Raftery, 2000).
- A default usage of conventional prior is not adaptive and may have a potential impact on the Bayes factors. Most of the model selection priors in the literature would have a similar misbehavior as they depend (implicitly) on the values of covariates (eg. through standardization).
- Using $\mathbf{\Sigma}^{uncens}$ "is not" allowed as it contains sample information (and does not contain information from censored data).

Observations

- The previous example is an extreme situation, but it is not unusual at all that $Var(X_{all}) >> Var(X_{uncens})$ (eg. the famous CHS and PBC survival datasets in Volinsky and Raftery, 2000).
- A default usage of conventional prior is not adaptive and may have a potential impact on the Bayes factors. Most of the model selection priors in the literature would have a similar misbehavior as they depend (implicitly) on the values of covariates (eg. through standardization).
- Using $\boldsymbol{\Sigma}^{uncens}$ "is not" allowed as it contains sample information (and does not contain information from censored data).
- Our prior should be able to adapt to situations with varying information content among units as is done by the likelihood. We derive such possibility using the expected information matrix.

Construction of a prior covariance matrix

Our plan is using:

$$\mathbf{\Sigma} = \Big(\text{Unitary information for } \boldsymbol{\beta}\Big)^{-1} = \text{eff. sample size} \times \big(\text{information for } \boldsymbol{\beta}\big)^{-1}$$

Construction of a prior covariance matrix

Our plan is using:

$$\mathbf{\Sigma} = \Big(\text{Unitary information for } \boldsymbol{\beta}\Big)^{-1} = \text{eff. sample size} \times \left(\text{information for } \boldsymbol{\beta}\right)^{-1}$$

- For (information for $\boldsymbol{\beta}$)$^{-1}$:
  - We use the block for $\boldsymbol{\beta}$ of the inverse of the expected information matrix. It has a closed-form expression but depends on $\boldsymbol{\beta}$ (a difficulty that does not appear in the linear model without censoring),

Construction of a prior covariance matrix

Our plan is using:

$$\boldsymbol{\Sigma} = \Big(\text{Unitary information for } \boldsymbol{\beta}\Big)^{-1} = \text{eff. sample size} \times \left(\text{information for } \boldsymbol{\beta}\right)^{-1}$$

- For $(\text{information for } \boldsymbol{\beta})^{-1}$:
  - We use the block for $\boldsymbol{\beta}$ of the inverse of the expected information matrix. It has a closed-form expression but depends on $\boldsymbol{\beta}$ (a difficulty that does not appear in the linear model without censoring),
  - We overcome that difficulty using *a la* Jeffreys trick: $\boldsymbol{\beta} = \mathbf{0}$

Construction of a prior covariance matrix

Our plan is using:

$$\boldsymbol{\Sigma} = \Big(\text{Unitary information for } \boldsymbol{\beta}\Big)^{-1} = \text{eff. sample size} \times \big(\text{information for } \boldsymbol{\beta}\big)^{-1}$$

- For (information for $\boldsymbol{\beta}$)$^{-1}$:
  - We use the block for $\boldsymbol{\beta}$ of the inverse of the expected information matrix. It has a closed-form expression but depends on $\boldsymbol{\beta}$ (a difficulty that does not appear in the linear model without censoring),
  - We overcome that difficulty using *a la* Jeffreys trick: $\boldsymbol{\beta} = \mathbf{0}$

- For effective sample size we borrow ideas from Berger et al. (2014), and use the expected information for β_0 in the null model (which in the linear model gets to n), leading to:

Construction of a prior covariance matrix

Our plan is using:

$$\mathbf{\Sigma} = \Big(\text{Unitary information for } \boldsymbol{\beta}\Big)^{-1} = \text{eff. sample size} \times \big(\text{information for } \boldsymbol{\beta}\big)^{-1}$$

- For (information for $\boldsymbol{\beta}$)$^{-1}$:
  - We use the block for $\boldsymbol{\beta}$ of the inverse of the expected information matrix. It has a closed-form expression but depends on $\boldsymbol{\beta}$ (a difficulty that does not appear in the linear model without censoring),
  - We overcome that difficulty using *a la* Jeffreys trick: $\boldsymbol{\beta} = \mathbf{0}$

• For effective sample size we borrow ideas from Berger et al. (2014), and use the expected information for β_0 in the null model (which in the linear model gets to n), leading to:

N

Construction of a prior covariance matrix

Our plan is using:

$$\boldsymbol{\Sigma} = \Big(\text{Unitary information for } \boldsymbol{\beta}\Big)^{-1} = \text{eff. sample size} \times \big(\text{information for } \boldsymbol{\beta}\big)^{-1}$$

- For (information for $\boldsymbol{\beta}$)$^{-1}$:
  - We use the block for $\boldsymbol{\beta}$ of the inverse of the expected information matrix. It has a closed-form expression but depends on $\boldsymbol{\beta}$ (a difficulty that does not appear in the linear model without censoring),
  - We overcome that difficulty using *a la* Jeffreys trick: $\boldsymbol{\beta} = \mathbf{0}$

• For effective sample size we borrow ideas from Berger et al. (2014), and use the expected information for β_0 in the null model (which in the linear model gets to n), leading to:

$$N_{(\beta_0,\sigma)} =$$

Construction of a prior covariance matrix

Our plan is using:

$$\boldsymbol{\Sigma} = \Big(\text{Unitary information for } \boldsymbol{\beta}\Big)^{-1} = \text{eff. sample size} \times \left(\text{information for } \boldsymbol{\beta}\right)^{-1}$$

- For (information for $\boldsymbol{\beta}$)$^{-1}$:
  - We use the block for $\boldsymbol{\beta}$ of the inverse of the expected information matrix. It has a closed-form expression but depends on $\boldsymbol{\beta}$ (a difficulty that does not appear in the linear model without censoring),
  - We overcome that difficulty using *a la* Jeffreys trick: $\boldsymbol{\beta} = \mathbf{0}$

• For effective sample size we borrow ideas from Berger et al. (2014), and use the expected information for β_0 in the null model (which in the linear model gets to n), leading to:

$$N_{(\beta_0,\sigma)} = \sum_{i=1}^{n} \omega_i, \quad \omega_i = \omega\left(\frac{c_i - \beta_0}{\sigma}\right), \quad \omega(z) = \Phi(z) + \phi(z)\left(\frac{\phi(z)}{1-\Phi(z)} - z\right).$$

...the effective sample size depends on $\boldsymbol{c}, \beta_0, \sigma$ and is unknown a priori!

Properties: about ω_i and effective sample size

$$N_{(\beta_0,\sigma)} = \sum_{i=1}^{n} \omega_i, \quad \omega_i = \omega\left(\frac{c_i - \beta_0}{\sigma}\right), \quad \omega(z) = \Phi(z) + \phi(z)\left(\frac{\phi(z)}{1-\Phi(z)} - z\right).$$

.

About $\omega(\cdot)$ and $N(\beta_0, \sigma)$:

- $0 \leq \omega(z) \leq 1$ and $\omega(z)$ increases with z,

Properties: about ω_i and effective sample size

$$N_{(\beta_0,\sigma)} = \sum_{i=1}^{n} \omega_i, \quad \omega_i = \omega\left(\frac{c_i - \beta_0}{\sigma}\right), \quad \omega(z) = \Phi(z) + \phi(z)\left(\frac{\phi(z)}{1-\Phi(z)} - z\right).$$

.

About $\omega(\cdot)$ and $N(\beta_0, \sigma)$:

- $0 \leq \omega(z) \leq 1$ and $\omega(z)$ increases with z,
- $0 \leq N(\beta_0, \sigma) \leq n$ and for fixed (β_0, σ):

Properties: about ω_i and effective sample size

$$N_{(\beta_0,\sigma)} = \sum_{i=1}^{n} \omega_i, \quad \omega_i = \omega\left(\frac{c_i - \beta_0}{\sigma}\right), \quad \omega(z) = \Phi(z) + \phi(z)\left(\frac{\phi(z)}{1-\Phi(z)} - z\right).$$

.

About $\omega(\cdot)$ and $N(\beta_0, \sigma)$:

- $0 \leq \omega(z) \leq 1$ and $\omega(z)$ increases with z,
- $0 \leq N(\beta_0, \sigma) \leq n$ and for fixed (β_0, σ):
  - $N_{(\beta_0,\sigma)} \to n$ if $c_i \to \infty, \forall i$,

Properties: about ω_i and effective sample size

$$N_{(\beta_0,\sigma)} = \sum_{i=1}^{n} \omega_i, \quad \omega_i = \omega\left(\frac{c_i - \beta_0}{\sigma}\right), \quad \omega(z) = \Phi(z) + \phi(z)\left(\frac{\phi(z)}{1-\Phi(z)} - z\right).$$

.

About $\omega(\cdot)$ and $N(\beta_0, \sigma)$:

- $0 \leq \omega(z) \leq 1$ and $\omega(z)$ increases with z,
- $0 \leq N(\beta_0, \sigma) \leq n$ and for fixed (β_0, σ):
  - $N_{(\beta_0,\sigma)} \to n$ if $c_i \to \infty, \forall i$,
  - $N_{(\beta_0,\sigma)} \to 0$ if $c_i \to -\infty, \forall i$,

Properties: about ω_i and effective sample size

$$N_{(\beta_0,\sigma)} = \sum_{i=1}^{n} \omega_i, \quad \omega_i = \omega\left(\frac{c_i - \beta_0}{\sigma}\right), \quad \omega(z) = \Phi(z) + \phi(z)\left(\frac{\phi(z)}{1-\Phi(z)} - z\right).$$

.

About $\omega(\cdot)$ and $N(\beta_0, \sigma)$:

- $0 \leq \omega(z) \leq 1$ and $\omega(z)$ increases with z,
- $0 \leq N(\beta_0, \sigma) \leq n$ and for fixed (β_0, σ):
  - $N_{(\beta_0,\sigma)} \to n$ if $c_i \to \infty, \forall i$,
  - $N_{(\beta_0,\sigma)} \to 0$ if $c_i \to -\infty, \forall i$,
  - $N_{(\beta_0,\sigma)} \to n_1$ if $\boldsymbol{c} = (c_u, \overset{n_1)}{\ldots}, c_u, c_c, \overset{n_2)}{\ldots}, c_c)$, and $c_c \to -\infty$.

Properties: about variance matrix

The covariance matrix adopts an appealing expression: it is a weighted covariance matrix:

$$\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma) = \sigma^2\left(\sum_{i=1}^{n}\omega_i(\boldsymbol{x}_i-\boldsymbol{x}_w)(\boldsymbol{x}_i-\boldsymbol{x}_w)^{\top}/N_{(\beta_0,\sigma)}\right)^{-1}, \;\; \boldsymbol{x}_w = \sum_{i=1}^{n}\omega_i\boldsymbol{x}_i/N_{(\beta_0,\sigma)},$$

that **Mixes** units using weights ω_i.

About $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma)$:

- $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma) = \boldsymbol{\Sigma}^{All}$, if all c_i are equal,

Properties: about variance matrix

The covariance matrix adopts an appealing expression: it is a weighted covariance matrix:

$$\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma)=\sigma^2\Big(\sum_{i=1}^{n}\omega_i(\boldsymbol{x}_i-\boldsymbol{x}_w)(\boldsymbol{x}_i-\boldsymbol{x}_w)^{\mathsf{T}}/N_{(\beta_0,\sigma)}\Big)^{-1},\ \ \boldsymbol{x}_w=\sum_{i=1}^{n}\omega_i\boldsymbol{x}_i/N_{(\beta_0,\sigma)},$$

that **Mixes** units using weights ω_i.

About $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma)$:

- $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma)=\boldsymbol{\Sigma}^{All}$, if all c_i are equal,
- If $\boldsymbol{c}=(c_u,\overset{n_1)}{\ldots},c_u,c_c,\overset{n_2)}{\ldots},c_c)$, and $c_c\to-\infty$, then $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma)\to\boldsymbol{\Sigma}^1$

Properties: about variance matrix

The covariance matrix adopts an appealing expression: it is a weighted covariance matrix:

$$\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma) = \sigma^2\Big(\sum_{i=1}^{n}\omega_i(\boldsymbol{x}_i-\boldsymbol{x}_w)(\boldsymbol{x}_i-\boldsymbol{x}_w)^{\top}/N_{(\beta_0,\sigma)}\Big)^{-1}, \;\; \boldsymbol{x}_w = \sum_{i=1}^{n}\omega_i\boldsymbol{x}_i/N_{(\beta_0,\sigma)},$$

that **Mixes** units using weights ω_i.

About $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma)$:

- $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma) = \boldsymbol{\Sigma}^{All}$, if all c_i are equal,
- If $\boldsymbol{c} = (c_u, \overset{n_1)}{\ldots}, c_u, c_c, \overset{n_2)}{\ldots}, c_c)$, and $c_c \to -\infty$, then $\boldsymbol{\Sigma}^{Mix}(\beta_0,\sigma) \to \boldsymbol{\Sigma}^{1}$

The resulting prior leads to finite marginals if $n_u \geq k+2$.

Predictive matching

General idea

When the sample is of minimal size, $n^\star$, then we should get a Bayes factor of 1 (exact predictive matching).

Predictive matching

General idea

When the sample is of minimal size, $n^\star$, then we should get a Bayes factor of 1 (exact predictive matching).

Bayarri et al. (2012) define several types of predictive matching criteria. The one that better characterizes aspects of the prior is:

Null predictive matching

Model selection priors are null predictive matching if $\{\mathcal{M}, \pi\}$ and $\{\mathcal{M}_0, \pi_0\}$ are exact predictive matching for samples of minimal size for $\mathcal{M}$.

Predictive matching

General idea

When the sample is of minimal size, $n^\star$, then we should get a Bayes factor of 1 (exact predictive matching).

Bayarri et al. (2012) define several types of predictive matching criteria. The one that better characterizes aspects of the prior is:

Null predictive matching

Model selection priors are null predictive matching if $\{\mathcal{M}, \pi\}$ and $\{\mathcal{M}_0, \pi_0\}$ are exact predictive matching for samples of minimal size for $\mathcal{M}$.

- In the linear model without censoring, Bayarri et al. (2012) show that for $n^\star = k + 1$, the priors:

$$\pi_0(\beta_0, \sigma) = \sigma^{-1}, \text{ and } \pi(\boldsymbol{\beta}, \beta_0, \sigma) = \sigma^{-1} \times \int N(\boldsymbol{\beta} \mid 0, g\boldsymbol{\Sigma})\pi(g)dg,$$

are exact predictive matching if and only if $\boldsymbol{\Sigma} = n\sigma^2(\widetilde{X}^T\widetilde{X})^{-1}$ (or proportional).

Censored data: samples of “minimal size” revisited and predictive matching

- Scenario I: $n^\star = k + 1$ [$n_c^\star \geq 1$, $n_u^\star \geq 2$ (for $m_0(\boldsymbol{y}, \boldsymbol{\delta})$ to exist)].

Censored data: samples of "minimal size" revisited and predictive matching

- Scenario I: $n^\star = k + 1$ [$n_c^\star \geq 1$, $n_u^\star \geq 2$ (for $m_0(\boldsymbol{y}, \boldsymbol{\delta})$ to exist)].

Result for Scenario I

$\boldsymbol{\Sigma}$ (known) leads to null predictive matching if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{All}$ (or a multiple).

Censored data: samples of "minimal size" revisited and predictive matching

- Scenario I: $n^\star = k+1$ [$n_c^\star \geq 1$, $n_u^\star \geq 2$ (for $m_0(\boldsymbol{y}, \boldsymbol{\delta})$ to exist)].

Result for Scenario I

$\boldsymbol{\Sigma}$ (known) leads to null predictive matching if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{All}$ (or a multiple).

• The information content vanishes for units with very small censoring times. Hence, in what information respects, a sample of "minimal size" is also

Censored data: samples of “minimal size” revisited and predictive matching

- Scenario I: $n^\star = k + 1$ [$n_c^\star \geq 1$, $n_u^\star \geq 2$ (for $m_0(\boldsymbol{y}, \boldsymbol{\delta})$ to exist)].

Result for Scenario I

$\boldsymbol{\Sigma}$ (known) leads to null predictive matching if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{All}$ (or a multiple).

• The information content vanishes for units with very small censoring times. Hence, in what information respects, a sample of “minimal size” is also

- Scenario II: $n^\star = n_c^\star + n_u^\star$ [$n_u^\star = k + 1$, $n_c^\star \geq 1$ with censoring times $c_i \to -\infty$.]

Censored data: samples of "minimal size" revisited and predictive matching

- Scenario I: $n^\star = k+1$ [$n_c^\star \geq 1$, $n_u^\star \geq 2$ (for $m_0(\boldsymbol{y}, \boldsymbol{\delta})$ to exist)].

Result for Scenario I

$\boldsymbol{\Sigma}$ (known) leads to null predictive matching if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{All}$ (or a multiple).

• The information content vanishes for units with very small censoring times. Hence, in what information respects, a sample of "minimal size" is also

- Scenario II: $n^\star = n_c^\star + n_u^\star$ [$n_u^\star = k+1$, $n_c^\star \geq 1$ with censoring times $c_i \to -\infty$.]

Result for Scenario II

$\boldsymbol{\Sigma}$ known leads to (limiting) null predictive matching

$$\lim_{c_i \to -\infty} B(\boldsymbol{y}, \boldsymbol{\delta}) = 1,$$

if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{uncens}$ (or a multiple).

What do we learn?

- From a predictive matching perspective, using $\mathbf{\Sigma}^{All}$ (all units equally contribute to the covariance matrix) is optimal for Scenario I (regular case), but it could be a bad choice for Scenario II (varying information content).

What do we learn?

- From a predictive matching perspective, using $\mathbf{\Sigma}^{All}$ (all units equally contribute to the covariance matrix) is optimal for Scenario I (regular case), but it could be a bad choice for Scenario II (varying information content).
- Our proposed covariance matrix $\mathbf{\Sigma}^{Mix}(\beta_0, \sigma)$ is adaptive, having $\mathbf{\Sigma}^{All}$ and $\mathbf{\Sigma}^{uncens}$ as particular cases and we interpret it as being the optimal choice in general.

What do we learn?

- From a predictive matching perspective, using $\mathbf{\Sigma}^{All}$ (all units equally contribute to the covariance matrix) is optimal for Scenario I (regular case), but it could be a bad choice for Scenario II (varying information content).
- Our proposed covariance matrix $\mathbf{\Sigma}^{Mix}(\beta_0, \sigma)$ is adaptive, having $\mathbf{\Sigma}^{All}$ and $\mathbf{\Sigma}^{uncens}$ as particular cases and we interpret it as being the optimal choice in general.

A curiosity: this adaptive behaviour comes with the price of a covariance matrix dependent on (β_0, σ) and for which the predictive matching criterion is not directly applicable (marginal exists for $n_u \geq k+2$).

Breast cancer survival in Castellon (Spain)

- $n = 2116$ women diagnosed with breast cancer in the decade 2004-2013,
- $y_i = log(t_i)$, where t_i is time to death (years) since diagnosis, which is censored for women who survived after the closing date: December 31st, 2015.

Breast cancer survival in Castellon (Spain)

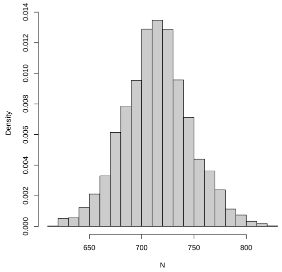
- $n = 2116$ women diagnosed with breast cancer in the decade 2004-2013,
- $y_i = log(t_i)$, where t_i is time to death (years) since diagnosis, which is censored for women who survived after the closing date: December 31st, 2015.
- Want to know evidence on the importance of $k = 6$ covariates ($2^6 = 64$ models): *number of nodes affected; age; recurrence (0/1); metastasis (0/1); estrogenic hormonal receptors (0/1) and progesterone hormonal receptors (0/1)*.

Breast cancer survival in Castellon (Spain)

- $n = 2116$ women diagnosed with breast cancer in the decade 2004-2013,
- $y_i = log(t_i)$, where t_i is time to death (years) since diagnosis, which is censored for women who survived after the closing date: December 31st, 2015.
- Want to know evidence on the importance of $k = 6$ covariates ($2^6 = 64$ models): *number of nodes affected; age; recurrence (0/1); metastasis (0/1); estrogenic hormonal receptors (0/1) and progesterone hormonal receptors (0/1)*.
- We observed $n_u = 360$ uncensored observations (83% of censoring).

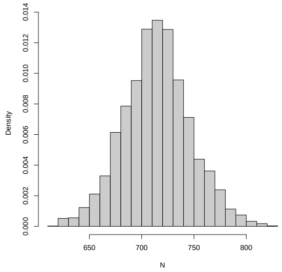
Breast cancer survival in Castellon (Spain)

- $n = 2116$ women diagnosed with breast cancer in the decade 2004-2013,
- $y_i = log(t_i)$, where t_i is time to death (years) since diagnosis, which is censored for women who survived after the closing date: December 31st, 2015.
- Want to know evidence on the importance of $k = 6$ covariates ($2^6 = 64$ models): *number of nodes affected; age; recurrence (0/1); metastasis (0/1); estrogenic hormonal receptors (0/1) and progesterone hormonal receptors (0/1).*
- We observed $n_u = 360$ uncensored observations (83% of censoring).



Breast cancer survival in Castellon (Spain)

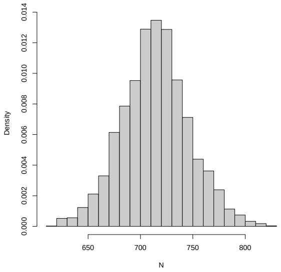
- $n = 2116$ women diagnosed with breast cancer in the decade 2004-2013,
- $y_i = log(t_i)$, where t_i is time to death (years) since diagnosis, which is censored for women who survived after the closing date: December 31st, 2015.
- Want to know evidence on the importance of $k = 6$ covariates ($2^6 = 64$ models): *number of nodes affected; age; recurrence (0/1); metastasis (0/1); estrogenic hormonal receptors (0/1) and progesterone hormonal receptors (0/1).*
- We observed $n_u = 360$ uncensored observations (83% of censoring).



- MA posterior distribution of the effective sample size $N_{(\beta_0,\sigma)}$ ($E(N \mid \text{data} = 714)$.

Breast cancer survival in Castellon (Spain)

- $n = 2116$ women diagnosed with breast cancer in the decade 2004-2013,
- $y_i = log(t_i)$, where t_i is time to death (years) since diagnosis, which is censored for women who survived after the closing date: December 31st, 2015.
- Want to know evidence on the importance of $k = 6$ covariates ($2^6 = 64$ models): *number of nodes affected; age; recurrence (0/1); metastasis (0/1); estrogenic hormonal receptors (0/1) and progesterone hormonal receptors (0/1)*.
- We observed $n_u = 360$ uncensored observations (83% of censoring).



- MA posterior distribution of the effective sample size $N_{(\beta_0,\sigma)}$ ($E(N \mid \text{data}) = 714$).

- Summary: $n_c = 1756$ 'count' as $E(N \mid \text{data}) - n_u = 354$ (20% information content)

Standard summaries of model selection based variable selection

{nodes, age, metasta, recurrence, ER, PGR}	0.473
{nodes, age, metasta, recurrence, ER}	0.467

Table: Posterior probabilities for the two most probable models.

Variable	nodes	age	metasta	recurrence	ER	PGR
Probability	1.00	1.00	1.00	0.98	0.96	0.52

Table: Breast cancer dataset: inclusion probabilities

Model averaging estimators

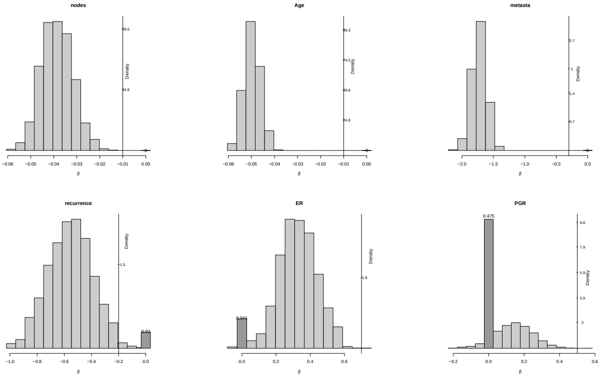


Figure: Breast cancer dataset. Model averaged posterior distributions of the regression coefficients for each potential covariate. Dark gray area represents the probability of no effect and the light gray area the distribution of probability given there is an effect.

Model averaging prediction (estimation of survival probabilities)

						Survival at year		
recurrence	metasta	nodes	age	ER	PGR	1	5	8
+	+	0	40	-	-	0.958	0.646	0.490
+	+	0	70	-	-	0.678	0.178	0.100
-	-	0	40	+	+	1	1	0.987
-	-	0	70	+	+	0.996	0.917	0.832
+	+	10	40	-	-	0.921	0.520	0.351
+	+	10	70	-	-	0.550	0.107	0.052
-	-	10	40	+	+	1	0.990	0.974
-	-	10	70	+	+	0.992	0.854	0.742
						0.999	0.941	0.873

Table: Last row is for an average case (values of the covariates at the sample mean).

Conclusions

- In general:
  - Use a Model Selection approach to variable selection, but with caution: Priors for model selection have particularities and, in general, shouldn't be blindly imported from similar problems.

Conclusions

- In general:
  - Use a Model Selection approach to variable selection, but with caution: Priors for model selection have particularities and, in general, shouldn't be blindly imported from similar problems.
  - A prior based on the expected information matrix seems to be a safe choice in general but it has to be scaled by a sensible effective sample size.

Conclusions

- In general:
  - Use a Model Selection approach to variable selection, but with caution: Priors for model selection have particularities and, in general, shouldn't be blindly imported from similar problems.
  - A prior based on the expected information matrix seems to be a safe choice in general but it has to be scaled by a sensible effective sample size.
  - This work is an exercise of how to construct such prior in problems with censored data...and we learn:

Conclusions

- In general:
  - Use a Model Selection approach to variable selection, but with caution: Priors for model selection have particularities and, in general, shouldn't be blindly imported from similar problems.
  - A prior based on the expected information matrix seems to be a safe choice in general but it has to be scaled by a sensible effective sample size.
  - This work is an exercise of how to construct such prior in problems with censored data...and we learn:
- For problems with censored data:
  - When Population$_{uncens}$ and Population$_{cens}$ are similar, then using Σ^{All} is a good choice because it is more precise than Σ^{uncens} (results will be more robust).

Conclusions

- In general:
  - Use a Model Selection approach to variable selection, but with caution: Priors for model selection have particularities and, in general, shouldn't be blindly imported from similar problems.
  - A prior based on the expected information matrix seems to be a safe choice in general but it has to be scaled by a sensible effective sample size.
  - This work is an exercise of how to construct such prior in problems with censored data...and we learn:
- For problems with censored data:
  - When $\text{Population}_{uncens}$ and Population_{cens} are similar, then using Σ^{All} is a good choice because it is more precise than Σ^{uncens} (results will be more robust).
  - When $\text{Population}_{uncens}$ and Population_{cens} differ, then using Σ^{All} is expected to produce more conservative results (than the preferred Σ^{uncens}).

Conclusions

- In general:
  - Use a Model Selection approach to variable selection, but with caution: Priors for model selection have particularities and, in general, shouldn't be blindly imported from similar problems.
  - A prior based on the expected information matrix seems to be a safe choice in general but it has to be scaled by a sensible effective sample size.
  - This work is an exercise of how to construct such prior in problems with censored data...and we learn:
- For problems with censored data:
  - When Population$_{uncens}$ and Population$_{cens}$ are similar, then using Σ^{All} is a good choice because it is more precise than Σ^{uncens} (results will be more robust).
  - When Population$_{uncens}$ and Population$_{cens}$ differ, then using Σ^{All} is expected to produce more conservative results (than the preferred Σ^{uncens}).
  - We do not know which situation is the real one, but our approach $\mathbf{\Sigma}^{Mix}$ provides a way to weight among these extreme possibilities, based on the censoring times.

Castilla-La Mancha

Figure: This work has been supported by grant SBPLY/17/180501/000491, funded by Consejería de Educación, Cultura y Deportes (JCCM, Spain) and FEDER and by Ministerio de Economa, Industria y Competitividad grant MTM2016-77501-P..

References I

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40:1550–1577.

Bayarri, M. J. and García-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, 94(1):135–152.

Berger, J., Bayarri, M., and Pericchi, L. (2014). The effective sample size. *Econometric Reviews*, 33(1-4):197–217.

Berger, J. O. and Pericchi, L. R. (2001). Objective bayesian methods for model selection: Introduction and comparison. In Lahiri, P., editor, *Model Selection*, volume 38, pages pp. 135–207. Institute of Mathematical Statistics.

Fernández, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100:381–427.

Held, L., Gravestocka, I., and Bové, D. S. (2016). Objective bayesian model selection for cox regression. *Statistics in medicine*, 35:5376–5390.

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, 3rd edition.

References II

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Nikooienejad, A., Wang, W., and Johnson, V. E. (2018). Bayesian variable selection for survival data using inverse moment priors. Technical Report arXiv:1712.02964, Cornell University.

Scott, J. and Berger, J. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

Sha, N., Tadesse, M., and Vanucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, 22(18):2262–2268.

Volinsky, C. T. and Raftery, A. E. (2000). Bayesian information criterion for censored survival models. *Biometrics*, 56(1):pp. 256–262.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Zellner, A., editor, *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, pages 389–399. Edward Elgar Publishing Limited.

Zellner, A. and Siow, A. (1980). Posterior odds ratio for selected regression hypotheses. In Bernardo, J. M., DeGroot, M., Lindley, D., and Smith, A. F. M., editors, *Bayesian Statistics 1*, pages 585–603. Valencia: University Press.