Bayesian Analysis (2021) **16**, Number 1, pp. 271–300

A Model Selection Approach for Variable Selection with Censored Data

María Eugenia Castellanos[*,§], Gonzalo García-Donato[†,§,¶], and Stefano Cabras[‡,§]

Abstract. We consider the variable selection problem when the response is subject to censoring. A main particularity of this context is that information content of sampled units varies depending on the censoring times. Our approach is based on model selection where all 2^k possible models are entertained and we adopt an objective Bayesian perspective where the choice of prior distributions is a delicate issue given the well-known sensitivity of Bayes factors to these prior inputs. We show that borrowing priors from the 'uncensored' literature may lead to unsatisfactory results as this default procedure implicitly assumes a uniform contribution of all units independently on their censoring times. In this paper, we develop specific methodology based on a generalization of the g-priors, explicitly addressing the particularities of survival problems arguing that it behaves comparatively better than standard approaches on the basis of arguments specific to variable selection problems (like e.g. predictive matching) in the particular case of the accelerated failure time model with lognormal errors. We apply the methodology to a recent large epidemiological study about breast cancer survival rates in Castellón, a province of Spain.

MSC2020 subject classifications: Primary 62C10, 62C10; secondary 62F15.

Keywords: Bayes factors, Bayesian model averaging, conventional priors, model selection, objective priors, predictive matching.

1 Introduction and motivation

In variable selection we have k possible explanatory variables but it is uncertain which of these is relevant to explain the response. We consider this situation in survival regression analysis where the response is subject to censoring.

Our research is rooted in the Bayesian paradigm and more concisely on methods based on the posterior distribution that assigns to each candidate model (a total of 2^k called the model space) its probability conditional on the observed data. The underlying general problem is normally called "model selection" on which, and unlike "estimation" problems, the true model is unknown. The posterior distribution is a simple function of the Bayes factors (Kass and Raftery, 1995) and the prior model probabilities. In this

[*]Universidad Rey Juan Carlos, Spain and Università degli Studi di Cagliari, Italy, maria.castellanos@urjc.es

[†]Universidad de Castilla-La Mancha, Spain, Gonzalo.GarciaDonato@uclm.es

[‡]Universidad Carlos III de Madrid, Spain, and Università degli Studi di Cagliari, Italy, stefano.cabras@uc3m.es

[§]Partially supported by the Ministerio de Ciencia e Innovación grants MTM2016-77501-P.

[¶]Partially supported by Junta de Comunidades de Castilla-La Mancha grant SBPLY/17/180501/000491/2.

 https://doi.org/10.1214/20-BA1207

paper we are mostly interested in objective methods, meaning that the prior inputs do not require any information *a priori*. See Berger (2006) and Consonni et al. (2018) for an updated review of objective Bayesian methods with particular emphasis on priors for model selection. The derivation of objective priors in model selection problems is very intriguing and a well known drawback of Bayes factors is a high sensitivity to prior specifications and that, in general, they cannot be derived using improper or vague priors (see e.g. Berger and Pericchi, 2001).

In survival problems, the information content in a sample varies among the different experimental units, defining likelihood functions that discriminate among censored and uncensored observations. In addition, objective priors are expected to reflect this particularity since these are normally derived as formal rules applied to the likelihood function (see Kass and Wasserman, 1996). In fact, in estimation problems, several authors have argued about the relevance of deriving specific objective priors in survival problems. For instance, De Santis et al. (2001) show that censoring considerations in the rules for deriving Jeffreys priors lead to proposals with better coverage frequentist intervals. In model selection problems, advances in Bayesian methods have mainly focused on relevant modeling aspects while the prior assignment is not specifically faced and, in general, these are borrowed from the uncensored literature. For instance Sha et al. (2006) develop variable selection methods in genetics using the original spike-and-slab priors by George and McCulloch (1993) and quite more recently (Nikooienejad et al., 2018), in a similar scenario but within Cox proportional hazards models, derive methods armed with the standard non-local priors by Johnson and Rossell (2010). Also Held et al. (2016) consider objective variable selection methods based on the test-based Bayes factors by Johnson (2008) implemented in combination with the original hyper g-priors (Liang et al., 2008) (with the interesting discussion on the convenience of using the number of uncensored observations and not the sample size to scale the prior covariance matrix).

Variable selection priors typically depend, in various ways, on the observed values of the covariates. This is the case of the g-prior, introduced for the linear model by Zellner and Siow (1980); Zellner (1986), and that is the basis for our proposal. For the normal regression model

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n),$$

where $\widetilde{\boldsymbol{X}}$ is the $n \times k$ design matrix with $\mathbf{1}^\top \widetilde{\boldsymbol{X}} = \mathbf{0}$ (so values of the covariates are centered around their mean), the g prior is a distribution for $\boldsymbol{\beta}$ conditional on (β_0, σ) that has the form:[1]

$$\boldsymbol{\beta} \mid \beta_0, \sigma \sim N_k(\mathbf{0}, \sigma^2 g\, n\, (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}). \tag{1.1}$$

The parameter g is either fixed at certain value (e.g. $g = 1$) or it is assumed to follow a distribution, $\pi(g)$, giving rise to the so called hyper-g priors (Zellner and Siow, 1980; Zellner, 1986; Liang et al., 2008; Bayarri et al., 2012). Clearly, the values of the covariates contribute to the g prior in the construction of the covariance matrix.

[1] In the literature it is more common to see the g prior parameterized in terms of gn so the covariance matrix is $\sigma^2 g\, (\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}})^{-1}$. Which representation is used is, of course, irrelevant and in this paper we have opted for the one introduced above because it better fits the arguments that accompany our proposal.

A main question that underlies our research is whether, in the situation with censored data, adopting the g prior (or more generally any 'uncensored' prior) without further considerations is a sensible approach. If this is done, then we are implicitly assuming that all values of covariates contribute uniformly to the prior even if they are associated with experimental units with very different censoring times. An illustration of the potential misbehavior of such default procedures is presented in Section 3 where, we show how a group of experimental units with very small censoring times (hence with negligible contribution to the likelihood) may severely modify the result of the variable selection exercise.

In this paper we derive generalizations of g and hyper-g priors for variable selection in survival problems when employing the accelerated failure time model with lognormal errors. The developed ideas are potentially useful for other type of parametric or semi-parametric models usually employed in survival analysis. This family of priors, that has been deeply studied in Berger and Pericchi (2001); Bayarri and García-Donato (2007) has received much attention in the literature and has been extended to problems beyond the original Gaussian model to include various types of error distributions (see e.g. Maruyama and Strawderman, 2010) but efforts have mainly focused on Generalized Linear Models (Sabanes and Held, 2011; Fouskakis et al., 2018; Li and Clyde, 2018). To the best of our knowledge this is the first attempt to extend g-priors to a context with censored observations.

Other full Bayesian possibilities to handle the variable selection problem are rules that, in principle, allow one to automatically obtain sensible priors. Among these, the most popular are those related to the intrinsic or fractional Bayes factors (O'Hagan, 1995; Berger and Pericchi, 1996; Moreno et al., 1998). These methods are strongly based on the concept of minimal training sample (see Berger and Pericchi, 2004, for a review of the topic), whose definition is particularly intriguing in problems with observations with different information content (as here). Strategies to circumvent these difficulties have been developed in the series of papers Perra et al. (2013); Cabras et al. (2014) and Cabras et al. (2015), but these approaches are computationally intensive since an integral must be evaluated for every training sample and many integrals are actually needed for one model comparison.

On the other hand, in the context of survival data, non Bayesian methods based on penalized likelihoods have been extended. In particular, Zhang and Lu (2007) proposed an adaptive LASSO for the Cox's proportional hazards model; Antoniadis et al. (2010) generalized to the case of the Cox proportional model the Dantzig selector methodology proposed by Candes and Tau (2007) in the regression linear model. In Fan and Li (2002) the smoothly clipped absolute deviation method (SCAD), a class of variable selection procedures using non-concave penalized likelihood, is used for variable selection also in the Cox model.

In the Bayesian setting, solving the variable selection problem reduces to finding ways to properly summarize the posterior distribution. To achieve this, specific summaries have been proposed, such as the model with the highest posterior probability or the marginal inclusion probabilities (Berger and Pericchi, 2001). Furthermore, Model Averaged estimates of any quantity of interest can be automatically derived, leading to

more realistic estimates (see Steel, 2019, and references therein, for a recent review of Model Averaging techniques). In our applications, we obtain such estimates to highlight their interest in applied problems.

The problem of selecting among two nested models contains the major challenges in a variable selection problem, but in terms of notation, it is much simpler to handle. With this in mind, we have structured the paper with the main theoretical sections (Sections 3 to 5) devoted to two model selection while the extension to variable selection is deferred until Section 6. Section 2 sets the notation for the Bayesian model selection problem in survival analysis, while Section 3 presents a motivating example. The main contents are presented in Section 4 where we construct our proposal; discuss several of its properties and introduce numerical approaches for its implementation. We illustrate the resulting approach to the problem with only two competing models in a classic transplant dataset in Section 2 of the supplementary material (Castellanos et al. 2020). In Section 5 we compare our proposal with the strategy of borrowing priors from the uncensored literature through the study of their predictive matching properties. In Section 7, we apply our method to analyse a recent epidemiological dataset on breast cancer in Spain. We further pursue the comparison among definitions of BF in Section 8 by means of an illustrative simulation study. Finally, some further remarks are offered in Section 9 while proofs of all results are provided in the supplementary material.

2 The statistical model considered

To introduce the statistical model, let y_i be the time-to-event (in logarithmic scale) for individual $i = 1, \ldots, n$ and assume that y_i follows a Gaussian distribution with density ϕ, cdf Φ and survival function $S = 1 - \Phi$. Due to censoring, the response is only observed if $y_i < c_i$, where c_i is (log) censoring time, and denote δ_i the binary variable that records a one if $y_i < c_i$ and zero otherwise. Once the experiment has finished, we observe which individuals have or have not been censored in the vector $\boldsymbol{\delta}^\intercal = (\delta_1, \ldots, \delta_n)$. For those $n_u = \sum_{i=1}^n \delta_i$ uncensored times, we observe their survival times $\boldsymbol{y}^\intercal = (y_1, \ldots, y_{n_u})$ (assuming without loss of generality that uncensored observations correspond to individuals $\{1, \ldots, n_u\}$). Likewise we denote $n_c = n - n_u$.

Throughout the paper we assume that (log) censoring times $\boldsymbol{c}^\intercal = (c_1, \ldots, c_n)$ are known, i.e. the closing time of the study is known. The use of this information to construct the prior will be a crucial differentiating aspect of the resulting methodology.

In this scenario, the joint density of $(\boldsymbol{y}, \boldsymbol{\delta})$ assuming independence among individuals, is

$$f(\boldsymbol{y}, \boldsymbol{\delta} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \phi\left(y_i \mid \boldsymbol{\theta}\right)^{\delta_i} \left[S\left(c_i \mid \boldsymbol{\theta}\right)\right]^{(1-\delta_i)}, \tag{2.1}$$

where $y_{n_u+1}, \ldots, y_n$ are immaterial as their corresponding $\delta_{n_u+1} = \ldots = \delta_n = 0$.

Suppose now that a set of k covariates labeled $x_1, \ldots, x_k$ are considered as potential explanatory variables and we want to test this possibility (model $\mathcal{M}_1$) against the model with just the intercept $\mathcal{M}_0$. Under $\mathcal{M}_1$

$$y_i = \beta_0 + \boldsymbol{\beta}^\intercal \widetilde{\boldsymbol{x}}_i + \sigma\varepsilon_i, \ \varepsilon_i \sim N(0, 1),$$

where $\widetilde{\boldsymbol{x}}_i$ is the k dimensional vector with the values of the covariates for individual i centered around the mean, that is:

$$\widetilde{\boldsymbol{x}}_i^{\mathsf{T}} = (x_{i1} - \bar{x}_1, x_{i2} - \bar{x}_2, \ldots, x_{ik} - \bar{x}_k),\ i = 1, \ldots, n.$$

Equivalently

$$\begin{aligned}\mathcal{M}_1 &: f_1(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) \\ &= \prod_{i=1}^{n} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - \beta_0 - \boldsymbol{\beta}^{\mathsf{T}}\widetilde{\boldsymbol{x}}_i}{\sigma}\right)\right]^{\delta_i} \left[1 - \Phi\left(\frac{c_i - \beta_0 - \boldsymbol{\beta}^{\mathsf{T}}\widetilde{\boldsymbol{x}}_i}{\sigma}\right)\right]^{(1-\delta_i)}, \end{aligned} \tag{2.2}$$

and the model with only the intercept:

$$\mathcal{M}_0 : f_0(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma) = \prod_{i=1}^{n} \left[\frac{1}{\sigma}\phi\left(\frac{y_i - \beta_0}{\sigma}\right)\right]^{\delta_i} \left[1 - \Phi\left(\frac{c_i - \beta_0}{\sigma}\right)\right]^{(1-\delta_i)}.$$

From these formulas it is easily deduced that censored observations with very negative c_i (recall c_i is in logarithmic scale), basically do not contribute to the likelihood.

The model $\mathcal{M}_0$ normally is called the null model while $\mathcal{M}_1$ is the full. In the literature, the parameters that appear in both models are called *common parameters*. With the centering performed to the covariates, the common parameter β_0 has a similar meaning in both the full and null models, representing the overall mean of survival times.

The full model can be written more compactly using matrix notation. For this, and as usual, denote $\widetilde{\boldsymbol{X}}$ the $n \times k$ matrix with $\widetilde{\boldsymbol{x}}_i^{\mathsf{T}}$ in its i-th row. Once the data is observed, this matrix can be partitioned (recall that we have assumed that uncensored observations occupy the first n_u positions) as $\widetilde{\boldsymbol{X}}^{\mathsf{T}} = (\widetilde{\boldsymbol{X}}_u^{\mathsf{T}}, \widetilde{\boldsymbol{X}}_c^{\mathsf{T}})$ where $\widetilde{\boldsymbol{X}}_u$ is $n_u \times k$ and $\widetilde{\boldsymbol{X}}_c$ is $n_c \times k$. This way:

$$f_1(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = N_{n_u}(\boldsymbol{y} \mid \mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_u\boldsymbol{\beta}, \sigma^2\boldsymbol{I})\, Pr\big(N_{n_c}(\mathbf{1}\beta_0 + \widetilde{\boldsymbol{X}}_c\boldsymbol{\beta}, \sigma^2\boldsymbol{I}) > \boldsymbol{c}_c\big), \tag{2.3}$$

where similarly to the notation above $\boldsymbol{c}^{\mathsf{T}} = (\boldsymbol{c}_u^{\mathsf{T}}, \boldsymbol{c}_c^{\mathsf{T}})$ and N_n represents a n-variate normal distribution with the first parameter being the vector mean and the second the covariance matrix.

Centered matrices relate with their non-centered counterparts $\boldsymbol{X}$, $\boldsymbol{X}_u$, $\boldsymbol{X}_c$ via the relation $\widetilde{\boldsymbol{X}} = (\boldsymbol{I} - n^{-1}\mathbf{1}_n\mathbf{1}_n^{\mathsf{T}})\boldsymbol{X}$, and

$$\widetilde{\boldsymbol{X}}_u = \boldsymbol{X}_u - \frac{1}{n}(\mathbf{1}_{n_u}\mathbf{1}_{n_u}^{\mathsf{T}}\boldsymbol{X}_u + \mathbf{1}_{n_u}\mathbf{1}_{n_c}^{\mathsf{T}}\boldsymbol{X}_c),\ \widetilde{\boldsymbol{X}}_c = \boldsymbol{X}_c - \frac{1}{n}(\mathbf{1}_{n_c}\mathbf{1}_{n_u}^{\mathsf{T}}\boldsymbol{X}_u + \mathbf{1}_{n_c}\mathbf{1}_{n_c}^{\mathsf{T}}\boldsymbol{X}_c), \tag{2.4}$$

where $\mathbf{1}_n$ represents the n-th dimensional column vector of ones (the subindex will be suppressed when possible).

i	y_i	δ_i	X_1	X_2	X_3	i	c_i	δ_i	X_1	X_2	X_3
1	−1.6	1	−0.4	0.4	0.6	7	−48	0	−0.1	6.5	−0.4
2	1.2	1	0.6	0.2	1.9	8	−48	0	1.1	9.3	−0.3
3	0.0	1	−0.1	−0.1	−0.6						
4	1.8	1	0.5	−0.0	−0.2						
5	1.3	1	−1.1	−1.0	−0.9	⋮	⋮	⋮	⋮	⋮	⋮
6	3.6	1	0.4	−0.6	0.6	30	−48	0	0.3	6.5	−0.6
$\bar{X}$			−0.02	−0.18	0.23				−0.05	5.6	−0.05

Table 1: Simulated dataset.

3 Motivating example

Consider the dataset in Table 1 with $n = 30$ of which, the first six observations are uncensored while the remaining 24 are censored with very small censoring times. Uncensored times are the log simulated values from an exponential model, with rate equal to the exponential of the linear predictor, following the scheme given in examples from the `R` package `BVSNLP` (Nikooienejad and Johnson, 2018). Three covariates ($k = 3$) have been used and parameters are fixed to $\beta_0 = -1$, $\beta_1 = -1.8$, $\beta_2 = 3$ and $\beta_3 = -0.7$, while X has been simulated from a multivariate normal with vector of means equal to 0, variances equal to 0.7 and covariances equal to 0.2. With respect to covariates, for censored observations, we change the mean of X_2 to be substantially larger, the means of X_1 and X_3 are similar to the ones of the uncensored observations. Particular values of the covariates are not important by themselves, so they have not been included in order to reduce the space. In this dataset note that the mean of X_2 substantially varies for the censored and uncensored observations.

What is relevant in this data set is that the contribution to the likelihood of the censored observations is negligible. In fact, a frequentist analysis based on lognormal regression, model in (2.2), produces exactly the same results (with very small p-values ($< 2 \times 10^{-16}$) for parameters β_1 and β_2 while the one associated with β_3 is 0.31) independently of whether the censored observations are used or not. We should expect a similar robustness to adding the censored observations in objective Bayesian implementations but this is not the case. For comparison, we have collected these results and the ones that follow in Table 2.

As fully described in Section 6, in the model selection approach to variable selection, the posterior probability of all models that arise as combinations of groups of the independent variables (including the null and the full model, hence a total of $2^3 = 8$ models in this problem) are obtained. These posterior probabilities are normally summarized with the posterior inclusion probabilities of the covariates (Barbieri and Berger, 2004) that contain the evidence in favor of a variable being truly explanatory for the response.

If we proceed this way in this data set, using the standard g prior given in (1.1) only considering the uncensored observations, that is with model (2.2) without the censoring part, we obtain inclusion probabilities (for x_1, x_2 and x_3) of 0.86, 0.90 and 0.3 (results obtained with the `R` package `BayesVarSel`, Garcia-Donato and Forte, 2018). On the other hand, if we consider the whole dataset, what means using model (2.2) with

the censored and observed data, and again the original g prior, (1.1),[2] the inclusion probabilities become 0.44, 0.51 and 0.22. The influence of the values of the covariates for the censored observations is remarkable.

This misbehavior is not exclusive of the g prior but of the strategy of directly importing priors from uncensored literature to survival problems. In this regard, and for illustrative purposes, we have used on this same dataset the methodology in Nikooienejad et al. (2018) (implemented with the accompanying `R` package `BVSNLP`) which develops the application of non-local priors originally proposed by Johnson and Rossell (2010) in survival models. The results of inclusion probabilities are 0.58, 0.78 and 0.24 just considering observations 1 to 6 and turn out to be 0.46, 0.46 and 0.16 respectively considering the whole data set.

This example is, purposely, quite extreme with a covariate (x_2) having quite a different distribution for censored and uncensored observations. Nevertheless, it warns about the need of considering generalizations of the standard objective priors for model selection specifically designed for survival problems. This is a main goal in this research which concludes in the objective prior that we propose in Section 4. Remarkably, when using this prior to the data set in Table 1 we obtain similar inclusion probabilities that were obtained with the g prior only considering the observations with non-negligible information content (the uncensored ones) and exactly the same results that using matrix Σ^U (4.14) as covariance matrix in (4.3).

4 Bayes factors, posterior probabilities and the prior proposed

Denoting $\pi_1(\beta_0, \sigma, \boldsymbol{\beta})$ the prior distribution for $\mathcal{M}_1$, the predictive density of $(\boldsymbol{y}, \boldsymbol{\delta})$, under $\mathcal{M}_1$, is

$$m_1(\boldsymbol{y}, \boldsymbol{\delta}) = \int f_1(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) \pi_1(\beta_0, \sigma, \boldsymbol{\beta}) d\beta_0 \, d\sigma \, d\boldsymbol{\beta}. \tag{4.1}$$

Similarly

$$m_0(\boldsymbol{y}, \boldsymbol{\delta}) = \int f_0(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma) \pi_0(\beta_0, \sigma) d\beta_0 \, d\sigma,$$

where $\pi_0(\beta_0, \sigma)$ denotes the prior under $\mathcal{M}_0$.

The Bayes factor for $\mathcal{M}_1$ against $\mathcal{M}_0$ and the posterior probability of $\mathcal{M}_1$ can be obtained, respectively as:

$$B_1(\boldsymbol{y}, \boldsymbol{\delta}) = \frac{m_1(\boldsymbol{y}, \boldsymbol{\delta})}{m_0(\boldsymbol{y}, \boldsymbol{\delta})}, \quad p(\mathcal{M}_1 \mid \boldsymbol{y}, \boldsymbol{\delta}) = \frac{B_1 \, p(\mathcal{M}_1)}{1 + B_1 \, p(\mathcal{M}_1)}, \tag{4.2}$$

where $p(\mathcal{M}_1)$ is the prior probability that $\mathcal{M}_1$ is the true model. When, as in this section, only two models are considered the objective choice is $p(\mathcal{M}_1) = 1/2$. The choice of the

[2] This corresponds to use matrix Σ^A in (4.6) as covariance matrix in (4.3).

	β_1	β_2	β_3
p-values:			
Model (2.2) with only uncensored data	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	0.31
Model (2.2) with whole dataset	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	0.31
Posterior inclusion probabilities based on model (2.2):			
Only uncensored data and g-prior, $g = 1$	0.86	0.90	0.3
Whole dataset and g-prior (1.1) with $g = 1$	0.44	0.51	0.22
Whole dataset with our proposal (4.3) with $g = 1$	0.85	0.88	0.29
Posterior inclusion probabilities based on Cox model:			
Only uncensored data	0.58	0.78	0.24
Whole dataset	0.46	0.46	0.16

Table 2: p-values and posterior inclusion probabilities (Bayesian analysis) associated with $\boldsymbol{\beta}$ parameters, when fitting the lognormal model or semiparametric Cox model (using the package `BVSNLP`) with different priors and either using only uncensored or the whole dataset in Table 1.

prior distribution in variable selection problems is less clear and we discuss about it in Section 6.

The prior for $\mathcal{M}_1$ can be written, $\pi_1(\boldsymbol{\beta}, \beta_0, \sigma) = \pi_1(\boldsymbol{\beta} \mid \beta_0, \sigma)\pi_1(\beta_0, \sigma)$. Our final proposal, as we largely discuss in the next subsections, is:

$$\pi_1(\boldsymbol{\beta} \mid \beta_0, \sigma) = \int N_k(\boldsymbol{\beta} \mid 0, g\,\boldsymbol{\Sigma})\,\pi(g)\,dg, \tag{4.3}$$

where

- $\pi(g)$ is defined in (4.5) (discussion in Section 4.1);
- $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^M$, a matrix defined in (4.11) (discussion in Section 4.2);
- Additionally, for the common parameters we use $\pi_1(\beta_0, \sigma) = \pi_0(\beta_0, \sigma) = \sigma^{-1}$ for the arguments introduced in Section 4.3.

Then, properties of the proposal are examined in Section 4.4. Finally Section 4.5 and Section 2 in the supplementary material are devoted to numerical implementation and a real illustration.

Later, in Section 5, it is argued that the proposal is nicely endorsed by predictive matching arguments.

4.1 General considerations

The form in (4.3) as a mixture of a multivariate normal distribution is ubiquitous in the literature of g priors (and extensions). Centering the prior at 0 was first proposed by Jeffreys (1961). This is a popular practice that has been formally justified using invariance arguments in Bayarri et al. (2012).

Above g is a mixing parameter that provides the prior with heavy tails through the mixing density $\pi(g)$, which give rise to several extensions of the g-prior in linear models. One of the first proposals is the Jeffreys-Zellner-Siow (JZS) prior (Jeffreys, 1961; Zellner and Siow, 1980), which implies using the inverse gamma

$$\pi(g) = \frac{1}{\sqrt{2\pi}} g^{-3/2} e^{-1/2g}, \quad g > 0. \tag{4.4}$$

More recently, Bayarri et al. (2012) have proposed the Robust prior

$$\pi(g) = \frac{1}{2}\sqrt{\frac{1+n}{n(k+1)}}\left(g + \frac{1}{n}\right)^{-3/2}, \quad g > \frac{1+n}{n(k+1)} - \frac{1}{n}, \tag{4.5}$$

that comes justified by a number of compelling arguments and criteria.

Other proposals assume a constant g, like the Zellner-g prior (Zellner, 1986; Kass and Wasserman, 1995) of using $g = 1$ (also called the Unit Information prior of Kass and Wasserman, 1995) or the Benchmark prior of Fernández et al. (2001) of using $g = \max\{1, k^2/n\}$.

Formulas for the model without censoring If all observations were uncensored (i.e. $\boldsymbol{\delta} = \mathbf{1}$) then the problem just presented exactly coincides with that of selecting among two Gaussian linear models, which has been examined in depth in the literature (see e.g. Bayarri et al. 2012 and references therein). As we mentioned in the introduction, the covariance matrix is normally chosen as

$$\boldsymbol{\Sigma}^A = n\sigma^2(\widetilde{\boldsymbol{X}}^\intercal \widetilde{\boldsymbol{X}})^{-1} \tag{4.6}$$

(the superindex A is used here to indicate that All observations are equally used). It is a well-known result that, with $\boldsymbol{\Sigma}^A$, the Bayes factor (again with $\boldsymbol{\delta} = \mathbf{1}$ so $\boldsymbol{y}$ is of dimension n) has the expression:

$$\mathcal{B}_\pi(\boldsymbol{y}, \widetilde{\boldsymbol{X}}, k, n) \equiv \int \left(1 + g\,\frac{SSE(\widetilde{\boldsymbol{X}})}{SSE_0}\right)^{-(n-1)/2} (1+g)^{(n-k-1)/2}\,\pi(g)\,dg, \tag{4.7}$$

with $SSE(\widetilde{\boldsymbol{X}}) = \boldsymbol{y}^\intercal(\boldsymbol{I} - \widetilde{\boldsymbol{X}}(\widetilde{\boldsymbol{X}}^\intercal\widetilde{\boldsymbol{X}})^{-1}\widetilde{\boldsymbol{X}}^\intercal)\boldsymbol{y}$, $SSE_0 = \boldsymbol{y}^\intercal(\boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal)\boldsymbol{y}$ (the sum of squared errors of the corresponding linear models). Notice that $\mathcal{B}_\pi(\boldsymbol{y}, \widetilde{\boldsymbol{X}}, k, n) = \mathcal{B}_\pi(\boldsymbol{y}, (\mathbf{1}\,\boldsymbol{X}), k, n)$ since $SSE(\widetilde{\boldsymbol{X}}) = SSE(\mathbf{1}\,\boldsymbol{X})$. This is a useful result that will be used later.

4.2 The prior covariance matrix

As early as in the seminal paper (Zellner and Siow, 1980) it has been recognized the importance of the Fisher information in providing a route to construct sensible prior covariance matrices for testing and model selection.

Consider the expected Fisher information matrix for regression parameters, $\mathcal{I}$, and its inverse, $\mathcal{J}$, being partitioned as

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{00} & \mathcal{I}_{01} \\ \mathcal{I}_{10} & \mathcal{I}_{11} \end{pmatrix}, \quad \mathcal{J} = \begin{pmatrix} \mathcal{J}_{00} & \mathcal{J}_{01} \\ \mathcal{J}_{10} & \mathcal{J}_{11} \end{pmatrix}.$$

Our first attempt would be using the covariance matrix equal to $\mathcal{J}_{11}$ above (without censoring this choice would lead to $\boldsymbol{\Sigma}^A$ in (4.6)). Despite the popularity of this matrix in the literature, its choice has been formally justified only recently. The justification was given in Bayarri et al. (2012) and is based on predictive matching criteria, and discriminates (4.6) with respect to other possibilities like $\mathcal{I}_{11}^{-1}$ or even assuming independence among parameters. In Section 5 we also utilize predictive matching arguments to justify our proposal. We derive $\mathcal{J}_{11}$ in the next result.

Theorem 4.1. *For the model defined in (2.2):*

- *The Hessian matrix corresponding to parameters $\boldsymbol{\beta}$ is:*

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} \log f_1(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}) = -\frac{1}{\sigma^2}\widetilde{\boldsymbol{X}}_u^\top \widetilde{\boldsymbol{X}}_u - \frac{1}{\sigma^2}\widetilde{\boldsymbol{X}}_c^\top \, Diag\{h(z_i)(h(z_i) - z_i)\}\, \widetilde{\boldsymbol{X}}_c,$$

  where $z_i = \frac{c_i - \beta_0 - \boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}_i}{\sigma}$, for $i = 1, \ldots, n$; $h(z) = \phi(z)/(1 - \Phi(z))$ is the hazard function and $Diag\{d_i\}$ stands for a diagonal matrix with values at the diagonal $d_1, \ldots, d_{n_c}$.

- *The expected Fisher information matrix corresponding to $(\beta_0, \boldsymbol{\beta})$ is*

$$\mathcal{I} = \frac{1}{\sigma^2}\begin{pmatrix} \mathbf{1}^\top \\ \widetilde{\boldsymbol{X}}^\top \end{pmatrix} \boldsymbol{W}(\beta_0, \boldsymbol{\beta}, \sigma) \begin{pmatrix} \mathbf{1} & \widetilde{\boldsymbol{X}} \end{pmatrix}, \tag{4.8}$$

  where $\boldsymbol{W}(\beta_0, \boldsymbol{\beta}, \sigma) = Diag\{\Phi(z_i) + \phi(z_i)(h(z_i) - z_i)\}$.

- *The block corresponding to $\boldsymbol{\beta}$ in the inverse matrix of $\mathcal{I}$ is*

$$\mathcal{J}_{11} = \sigma^2 \Big(\widetilde{\boldsymbol{X}}^\top \Big[\boldsymbol{W}(\beta_0, \boldsymbol{\beta}, \sigma) - \boldsymbol{W}(\beta_0, \boldsymbol{\beta}, \sigma) \frac{\mathbf{1}\mathbf{1}^\top}{tr\boldsymbol{W}(\beta_0, \boldsymbol{\beta}, \sigma)} \boldsymbol{W}(\beta_0, \boldsymbol{\beta}, \sigma) \Big] \widetilde{\boldsymbol{X}} \Big)^{-1}, \tag{4.9}$$

  where $tr\boldsymbol{A}$ stands for the trace of $\boldsymbol{A}$.

Unfortunately, the matrix in (4.9) is useless for our purposes as it depends on the parameter $\boldsymbol{\beta}$ for which we are constructing the prior distribution. This is a differentiating aspect of models outside the linear model under which this matrix does not depend on $\boldsymbol{\beta}$. To overcome this difficulty we adopt the traditional idea of 'importing' the assessment under the null model (i.e. $\boldsymbol{\beta} = \mathbf{0}$) with the underlying idea that, in the variable selection problem, the only prior available information about the non-common parameters is that assessed by the null model that states that, with positive probability, these parameters are zero. Another very interesting alternative would be using the MLE of $\boldsymbol{\beta}$ to surpass

the dependence of the matrix on the unknown parameters. The resulting procedure, which has been used for instance by Clyde (1999) in GLM's, would have an empirical Bayes flavor and is undoubtedly an alternative to our 'null' based approach; however, to the greatest extent possible, we prefer using only genuine prior distributions.

Another source of concern regarding a direct use of $\mathcal{J}_{11}$ is that it is quite informative as it contains the information in the whole sample and if it is directly used, the prior will be very influential. Hence, we need to rescale this matrix by a certain *effective* sample size, here denoted as N, to capture the notion of a prior as informative as a sample of size one (for related ideas see Bayarri and García-Donato, 2008). In the great majority of occasions, for practical purposes (see e.g. Kass and Raftery, 1995) the effective sample size is simply taken as the sample size n. Nevertheless, our problem is a clear example where using n is not a sensible option since the information content in observational units depends on whether they are (or are not) censored. Intuitively, censored units should contribute less to the construction of the effective sample size. Unfortunately, a priori it is unknown which observations are censored and hence we need to rely on some sort of weight that gives more importance to units that, in a certain sense, are more expected to be uncensored. It is here where the use of the censoring times c_i will be of utmost importance.

There have been many attempts to define the notion of effective sample size and the most comprehensive study is the recent paper (Berger et al., 2014). The underlying idea is to choose the effective sample size roughly, like the precision.

One particular practical conclusion that can be extracted from Berger et al. (2014) is that in linear models, the effective sample size, defined as the expected information for the intercept in the null model, coincides with the sample size n. This also holds approximately true for regression models with no pathologic values of the explanatory variables.

In our setting, the information under the null can be easily obtained from (4.8) leading to

$$N_{(\beta_0,\sigma)} \equiv \sum_{i=1}^{n} \Phi(z_{i0}) + \phi(z_{i0})\Big(\frac{\phi(z_{i0})}{1-\Phi(z_{i0})} - z_{i0}\Big), \tag{4.10}$$

where $z_{i0} = (c_i - \beta_0)/\sigma$. In Section 2 in the supplementary material we show that the unknown parameter $N_{(\beta_0,\sigma)}$ captures quite precisely the idea of an effective sample size in a problem with censored observations.

With all the considerations above, our proposal for the covariance matrix is

$$\boldsymbol{\Sigma}^M = N_{(\beta_0,\sigma)}\sigma^2\big[\widetilde{\boldsymbol{X}}^{\intercal}\big(\boldsymbol{W}(\beta_0,\sigma) - \boldsymbol{W}(\beta_0,\sigma)\frac{\mathbf{1}\mathbf{1}^{\intercal}}{N_{(\beta_0,\sigma)}}\boldsymbol{W}(\beta_0,\sigma)\big)\widetilde{\boldsymbol{X}}\big]^{-1}, \tag{4.11}$$

with

$$\boldsymbol{W}(\beta_0,\sigma) = \boldsymbol{W}(\beta_0,\boldsymbol{\beta}=\mathbf{0},\sigma) = Diag\{w_i(\beta_0,\sigma)\},$$

and $w_i(\beta_0,\sigma) = w(z_{i0})$, for $i = 1,\ldots,n$, and

$$w(z) = \Phi(z) + \phi(z)(h(z) - z). \tag{4.12}$$

With this notation, $N_{(\beta_0,\sigma)} = \sum_{i=1}^{n} w_i(\beta_0, \sigma)$. The index M in (4.11) refers to the idea that a mix of censored and uncensored observations is used.

Note that the matrix in (4.11) depends on (β_0, σ) in a way that comes inspired by the null model, hence it is important that both parameters have 'the same meaning' throughout the different models. It is at this point (and only at this point) where introducing the model with covariates being centered around their means is very important. In fact, in this parametrization β_0 has the interpretation of being the conditional expected value of Y when covariates equals their respective means, so that if covariates have the same mean (i.e. zero) also β_0 has the same interpretation independently of the assumed model.

In what follows, and unless needed, the dependence on (β_0, σ) of N, $\boldsymbol{W}$ and w_i will be removed to simplify the notation.

4.3 About the prior over common parameters

Our proposal for the prior for common parameters is the standard location-scale invariant prior $\pi(\beta_0, \sigma) = \sigma^{-1}$. Another possibility would be to include the idea of censoring also in this prior (e.g. the Jeffreys priors).

Considering the likelihood for the null model, the expression of the Fisher Information matrix $\mathcal{I}(\beta_0, \sigma)$ is given in Equation (2) in the proof of Theorem 1 in the supplementary material. This result clearly suggests that Jeffreys prior using the full Fisher information is computationally much more expensive than $\pi(\beta_0, \sigma) = \sigma^{-1}$.

Therefore, being not clear if the complication with respect to the standard option is really needed, we performed a comparison of the following two priors at hand:

1. $\pi^J(\beta_0, \sigma) = \sqrt{|\mathcal{I}(\beta_0, \sigma)|}$, based on Jeffreys prior using the full Fisher information matrix, derived above;
2. $\pi(\beta_0, \sigma) = 1/\sigma$, Jeffreys prior corresponding to the model without censoring.

While comparisons should be made more formally in terms of model selection performances, an analysis of coverages often suffices. We thus analyze the coverage in estimating the common parameters β_0, σ under the null model, excluding the other parameters (in other models). The results with regard to coverage of 500 nominal 95% credible intervals considering two different sample sizes, $n = 30$ and $n = 50$ and four proportions of censoring, 0.2, 0.4, 0.6 and 0.8 appear in Figure 1.

We can see that almost all coverages reported in Figure 1 are compatible among the two priors, and near of the nominal 95% considering their respective standard errors, except for both priors at low sample sizes or data with a very high proportion of censored observations. However, low sample size and very high proportion of censored observations is a somewhat unrealistic scenario at least compared to the considered datasets. These results suggest that we should not expect large differences between using either prior and here we adopt the more convenient choice (from a computational point of view) of using $\pi(\beta_0, \sigma) \propto 1/\sigma$.

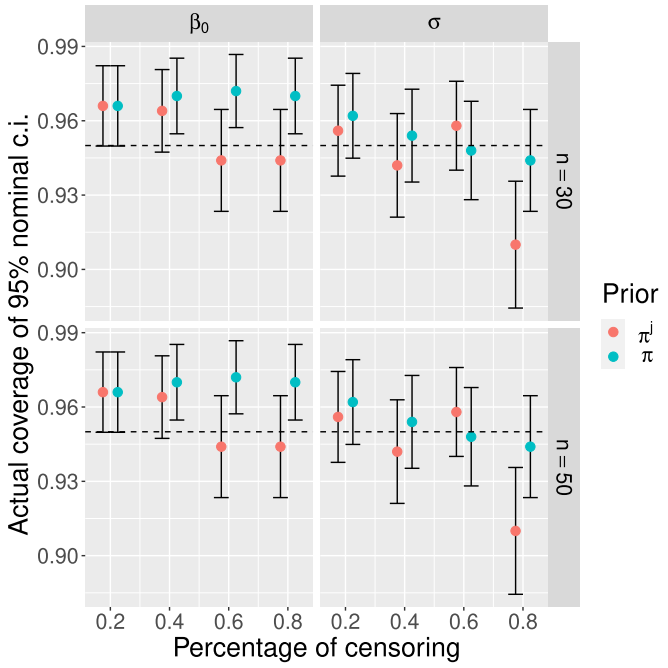

Figure 1: Actual coverages of 95% credible intervals for parameters β_0 and σ under the two priors $\pi^J(\beta_0, \sigma) = \sqrt{|\mathcal{I}(\beta_0, \sigma)|}$ and $\pi(\beta_0, \sigma) = 1/\sigma$ (intervals are obtained with $\pm 2\times$ standard error).

4.4 Properties of Σ^M and the proposed prior

In this section the properties and interpretation of the matrix (4.11) are examined. The main conclusion that we extract is that, based on the prior information in the censoring times c_i, this matrix has the ability to automatically adjust for the information content in the different observations.

Consider first the following result that concerns the function $w(z)$ in (4.12) and an equivalent expression for $\mathbf{\Sigma}^M$ in (4.11).

Theorem 4.2. *It holds true that*

i) $0 \leq w(z) \leq 1$ *for all* $z \in \mathcal{R}$.

ii) $w(z)$ *is an increasing function.*

iii)

$$\mathbf{\Sigma}^M = \sigma^2 \Big(\sum_{i=1}^{n} w_i (\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_w)(\widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_w)^\intercal / N \Big)^{-1}, \tag{4.13}$$

where $\widetilde{\boldsymbol{x}}_w = \sum_{i=1}^{n} w_i \widetilde{\boldsymbol{x}}_i / N$.

From (4.13) it is easy to deduce that *given* (β_0, σ) the matrix $\mathbf{\Sigma}^M$ is equivariant to changes in location of the covariates, and in particular

$$\mathbf{\Sigma}^M = \sigma^2 \Big(\sum_{i=1}^{n} w_i (\boldsymbol{x}_i - \boldsymbol{x}_w)(\boldsymbol{x}_i - \boldsymbol{x}_w)^\intercal / N_{(\beta_0,\sigma)} \Big)^{-1},$$

where $\boldsymbol{x}_w = \sum_{i=1}^{n} w_i \boldsymbol{x}_i / N_{(\beta_0,\sigma)}$.

Because of the positiveness of w_i, it is valid to interpret them as *weights* that, according to *ii*), are a monotone increasing function of $(c_i - \beta_0)/\sigma$. This implies that the larger the standardized difference $c_i - \beta_0$ is, (it is more likely that individual i is uncensored), the larger the value of the weight. In the extremes, if $c_i \to \infty$ then w_i tends to one and if $c_i \to -\infty$ then $w_i \to 0$. This fact plus the equivalence *iii*) justifies that the proposed prior covariance matrix is proportional to a weighted covariance matrix of the explanatory variables where those units i with larger c_i contribute more to the covariance matrix.

The w_i also provide an interesting interpretation of the effective sample size N (which we recall is defined as $\sum_{i=1}^{n} w_i$). Clearly $0 < N \leq n$ and $N \to n$ if all $c_i \to \infty$. In the other extreme, if all $c_i \to -\infty$ then $N \to 0$. Hence the proposed effective sample size is a compromise between the number of units that are expected to be uncensored (large c_i) and censored (small c_i).

Note that these results (and many that follow) hold true in the limit (e.g. when c tends either to ∞ or $-\infty$ because $w((c - \beta_0)/\sigma)$ tends to 1 or 0) but, for practical purposes, it is important to know that up to the second decimal point $w(z) \approx 1$ if $z > 1.50$ (since $w(1.5) = 0.9900$) while $w(z) \approx 0$ when $z < -3.15$ ($w(-3.15) = 0.0096$). This means that the results when $w((c-\beta_0)/\sigma) \to 1$ are essentially true if c is moderately large compared with β_0 (if c is farther than 1.5 standard deviations from β_0). Similarly, $w((c - \beta_0)/\sigma) \approx 0$ if $c < \beta_0 - 3.15\sigma$.

There are two specific scenarios where the expression adopted by $\mathbf{\Sigma}^M$ is particularly revealing and that we now analyze.

Homogeneous censoring times When all c_i coincide, and without using any external information, it is virtually impossible to value, a priori, which units are more likely to be censored. Our proposed approach agrees with this observation and it can be easily seen that $\mathbf{\Sigma}^M$ would coincide with $\mathbf{\Sigma}^A$ (cf. (4.6)), the equally weighted sample covariance matrix of regressors and the prior covariance matrix used in the g-prior in the problem without censoring.

Polarized censoring times An opposite situation to that just considered is when the sample units are highly polarized with some of them having very small censoring times compared with the rest. A canonical of such cases is when

$$\boldsymbol{c} = (c_u, \overset{n_u}{\ldots}, c_u, c_c, \overset{n_c}{\ldots}, c_c),$$

and c_c tends to be arbitrarily small. Here, and for fixed (β_0, σ), $\boldsymbol{\Sigma}^M \to \boldsymbol{\Sigma}^U$, the covariance matrix only based on uncensored observations, that is,

$$\boldsymbol{\Sigma}^U = n_u \sigma^2 (\boldsymbol{X}_u^\top (\boldsymbol{I} - \mathbf{1}_{n_u} \mathbf{1}_{n_u}^\top / n_u) \boldsymbol{X}_u)^{-1}. \tag{4.14}$$

This implies that the proposed prior would coincide with the g-prior considering only the values of the covariates from the uncensored individuals. As we demonstrate in the next result, under these conditions and, as expected (since the contribution of last n_c observations to likelihood tends to vanish), the Bayes factor would tend to the Bayes factor that only takes into consideration the uncensored observations.

Theorem 4.3. *Suppose the hyper parameter g is fixed or $\int_0^\infty g^{-k/2}\, \pi(g)\, dg < \infty$. Let $\boldsymbol{c}^\top = (c_u, \overset{n_u}{\ldots}, c_u, c_c, \overset{n_c}{\ldots}, c_c)$ and let the matrix $\boldsymbol{X}$ be partitioned accordingly as $\boldsymbol{X}^\top = (\boldsymbol{X}_u^\top, \boldsymbol{X}_c^\top)$ with conformable dimensions. If $n_u \geq k+2$ then*

$$\lim_{c_c \to -\infty} B_1(\boldsymbol{y}, \boldsymbol{\delta}) = \mathcal{B}_\pi(\boldsymbol{y}, (\mathbf{1}\, \boldsymbol{X}_u), k, n_u).$$

Note that it can be easily seen that the condition $\int_0^\infty g^{-k/2}\, \pi(g)\, dg < \infty$ is satisfied for the priors (4.4) and (4.5). This result will become more relevant in a later discussion about predictive matching properties.

Arguably these last two scenarios are extreme and of a limited utility in practice. Nevertheless, they illustrate how our proposed Bayes factor is somewhere in between (averaged by the information in c_i) the two alternative approaches of considering all the observations in the same manner to construct the prior or just considering the censored observations.

The dependence of (4.11) on (β_0, σ) implies a substantial change both from a theoretical and a numerical perspective (we discuss the latter in Section 5). Hence, a crucial question is to know if this dependence changes the conditions for the existence of the conventional Bayes factor for the problem without censoring, where it is known (see e.g. Bayarri et al., 2012) that a sufficient and necessary condition is $n \geq k+1$ (it is necessary to ensure the existence of the inverse matrix (4.6)).

For the proposed prior, in the next result we show that, for the case where all c_i are equal, then $n \geq k+1$ when $n_u \geq 2$ (recall n is the number of observations, censored or uncensored) still ensures propriety of posterior. Unfortunately, the problem is more involved when censoring times are not constant, precisely due to the real dependence on (β_0, σ). In this last case we show that $n_u \geq k+2$ is a sufficient condition for posterior propriety.

Theorem 4.4. *If either*

i) $n \geq k+1$, $n_u \geq 2$ *and* $c_i = c$ *for all* i, *or*

ii) $n_u \geq k+2$ *and* g *is constant or* $\int g^{-k/2}\, \pi(g)\, dg < \infty$,

then $0 < m_1(\boldsymbol{y}, \boldsymbol{\delta}) < \infty$ *(or equivalently the posterior is proper).*

Finally, the prior has the property of leading to Bayes factors that are equivariant to changes in scale and/or location of the explanatory variables.

Theorem 4.5. *Let the original design matrix* $\boldsymbol{X}$ *be transformed as* $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{D} + \mathbf{1}_n \boldsymbol{b}^\top$, *where* $\boldsymbol{D}$ *is a non-singular diagonal matrix of dimension* $k \times k$ *and* $\boldsymbol{b} \in \mathcal{R}^k$, *both with known entries. Then, the Bayes factor remains the same independently of whether* $\boldsymbol{X}$ *or* $\boldsymbol{Z}$ *are used in the construction of* $\boldsymbol{\Sigma}^M$ *in* (4.11).

4.5 Computing Bayes factors

The marginal predictive distribution for each model is not available in closed-form and has been approximated using importance sampling, in particular the algorithm proposed in Touloupou et al. (2018).

The steps we follow to obtain the approximation of each predictive distribution are:

1. Obtain a sample from the posterior distribution $\pi(\boldsymbol{\beta}, \beta_0, \sigma, g \mid \boldsymbol{y}, \boldsymbol{\delta})$, using Markov chain Monte Carlo (MCMC) methods. In particular, a random walk Metropolis-Hastings (RW-MH) algorithm has been employed initialized at the mode of the posterior distribution for $\boldsymbol{\beta}, \beta_0, \log(\sigma)$ and in $g = 1$. The RW-MH uses for parameters $(\boldsymbol{\beta}, \beta_0, \log(\sigma))$, a multivariate normal proposal having a variance proportional to the inverse of the Hessian at the mode previously obtained and with a fixed scale factor, chosen in order to have an acceptance rate of 30–40%. Such proposal's variance is calculated over a fraction of the burn-in period and then it stays steady for the rest of the steps. Independently for g we use as proposal the prior distribution (4.4) or (4.5). This algorithm has been also used in Cabras et al. (2014) and since the priors are proper, the convergence of the chain to the target posterior can be assumed. More details about the employed proposal distributions and an explicit expression of the target one are provided in the supplementary material.

2. Construct a proposal distribution, $q(\cdot)$, based on the posterior sample obtained in the above point. For $(\boldsymbol{\beta}, \beta_0, \log(\sigma))$ we have used a multivariate t-distribution with 3 degrees of freedom and noncentrality parameters and scale matrix constructed with the posterior sample previously obtained and for g we use as proposal distribution the prior one.

3. Sample $((\boldsymbol{\theta}_1, g_1), \ldots, (\boldsymbol{\theta}_N, g_N))$ from the proposal distribution $q(\cdot)$, and use the importance estimator to approximate the marginal predictive distribution; here we are denoting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \beta_0, \sigma)$:

$$\widehat{m(\boldsymbol{y}, \boldsymbol{\delta})} = \frac{1}{N} \sum_{i=1}^{N} \frac{\pi(\boldsymbol{\theta}_i, g_i \mid \boldsymbol{y}, \boldsymbol{\delta})}{q(\boldsymbol{\theta}_i, g_i)}.$$

4. The approximated Bayes factor is

$$\widehat{B_1(\boldsymbol{y}, \boldsymbol{\delta})} = \frac{\widehat{m_1(\boldsymbol{y}, \boldsymbol{\delta})}}{\widehat{m_0(\boldsymbol{y}, \boldsymbol{\delta})}}.$$

More details about the implementation are included in Section 3 in the supplementary material.

Of course, other ways of approximating Bayes factors from MCMC, such as bridge sampling (DiCiccio et al., 1997), are possible. However, they are outside the scope of the paper.

5 Predictive matching

Predictive matching (PM) is a powerful principle to judge the suitability of objective priors for comparing models of varying dimensions (see Bayarri et al., 2012). PM can be seen as a combination of two arguments: i) a convenient way to comparing priors is through their corresponding predictive distributions (Berger and Pericchi, 2001) and ii) one should not be able to choose among competing models when the sample information is extremely small (this argument was first suggested by Jeffreys, 1961). As a result, the PM principle states that two priors are properly calibrated for model comparison if their predictive distributions match when the sample is of minimal size.

We devote this section to study PM aspects of the priors proposed in Section 4. When relevant, these results are compared with the prior using equally all experimental units (so $\boldsymbol{\Sigma}^A$ in (4.6) is taken as prior covariance matrix) and when only the uncensored observations are considered ($\boldsymbol{\Sigma}^U$ in (4.14) is used).

It is convenient to first revisit the concept of predictive matching (PM) for the problem without censoring (so $n_u = n, n_c = 0$), where the idea was originally developed. In this context the underlying model is the standard linear model, $\boldsymbol{\delta} = 1$ and we simply write $m_i(\boldsymbol{y}, \boldsymbol{\delta}) = m_i(\boldsymbol{y})$. The intuition behind PM is that when the information in the sample is tiny, then one should not be able to reach a decision and so the Bayes factor should be one. The notion of being very scarce in information is usually associated with a sample $\boldsymbol{y}^\star$ of minimal size, meaning that $0 < m_i(\boldsymbol{y}^\star) < \infty$, for $i = 0, 1$, and for samples of smaller sizes the marginal is either zero or infinite. In Bayarri et al. (2012) several definitions of minimal size are considered depending on which type of prior is used in $m_i(\boldsymbol{y}^\star)$. Of these, the one that leads to a stronger requirement on the construction of priors and the one that we later adopt in our setting uses $k+1$ (the number of unknown parameters in the full model) as minimal size. For samples of this minimal size (Bayarri et al., 2012) show that having a covariance matrix of the form in (4.6) (or a multiple) is a necessary and sufficient condition for predictive matching. This result provides a formal justification for the use of this covariance prior matrix (the only result that exists with these characteristics to the best of our knowledge).

In what follows, and without loss of generality, we assume that the values in $\boldsymbol{y}$ are all different and that $n_c \geq 1$ (otherwise the results for the linear model without censoring in Bayarri et al., 2012, apply).

A necessary condition for Predictive Matching results to apply is that the marginal under the null exists. In our problem this function has the form:

$$m_0(\boldsymbol{y}, \boldsymbol{\delta}) = \int \prod_{i \in \boldsymbol{c}_c} \left(1 - \Phi\left(\frac{c_i - \beta_0}{\sigma}\right)\right) \times N_{n_u}(\boldsymbol{y} \mid \beta_0 \mathbf{1}, \sigma^2 \boldsymbol{I}) \frac{1}{\sigma} \, d\sigma d\beta_0.$$

In the next result we prove that $m_0(\boldsymbol{y}, \boldsymbol{\delta})$ exists if we have at least two uncensored observations.

Lemma 1. *The marginal $m_0(\boldsymbol{y}, \boldsymbol{\delta})$ is finite if and only if $n_u \geq 2$.*

Retaking the PM criterion, the relevant question now is which conditions must satisfy a prior of the form (4.3) if we have $k+1$ data points. This in principle leads us to the scenario where we have $n = k+1$ (at least two of these being uncensored for m_0 to exist). Nevertheless, the interesting characteristic in censored models is that censored observations do not contribute to the likelihood in a similar way as is done by uncensored data and, in fact, their contribution vanishes when $c_i \to -\infty$. Following this argument, a situation with $n_u = k+1$ and an arbitrary number n_c of censored observations for which $c_\star = \max\{c_i \in \boldsymbol{c}_c\} \to -\infty$ is clearly a scenario of minimal size. In summary, there are two different scenarios of minimal size:

- Scenario I: $n = k+1$, $n_c \geq 1$, $n_u \geq 2$,
- Scenario II: $n_u = k+1$, $n_c \geq 1$ and $c_\star \to -\infty$.

Note that the conditions defined above do not overlap and the number of uncensored observations in Scenario I varies in $2 \leq n_u \leq k$ while in Scenario II it is $n_u = k+1$. We derive two PM-type results that we later interpret.

Lemma 2. *Consider the comparison of $\mathcal{M}_1$ and $\mathcal{M}_0$ using the prior (4.3) with $\boldsymbol{\Sigma}$ independent of (β_0, σ) and $\pi_0(\beta_0, \sigma) = \pi_1(\beta_0, \sigma) = \sigma^{-1}$. Then, in the conditions of Scenario I exact predictive matching is attained (i.e. $B_1(\boldsymbol{y}, \boldsymbol{\delta}) = 1$) if and only if $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^A$ (or a multiple).*

Lemma 3. *Consider the comparison of $\mathcal{M}_1$ and $\mathcal{M}_0$ using the prior (4.3) with $\boldsymbol{\Sigma}$ independent of (β_0, σ) and $\pi_0(\beta_0, \sigma) = \pi_1(\beta_0, \sigma) = \sigma^{-1}$. Then, in the conditions of Scenario II limiting exact predictive matching is attained:*

$$\lim_{c_\star \to -\infty} B_1(\boldsymbol{y}, \boldsymbol{\delta}) = 1,$$

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

These two lemmas are somehow contradictory since both $\boldsymbol{\Sigma}^A$ and $\boldsymbol{\Sigma}^U$ (which cannot be used in a prior because it contains sample information) are endorsed by Predictive Matching arguments. Obviously these matrices can be very different particularly if covariates for censored and uncensored observations exhibit different behavior. Lemma 2 suggests that $\boldsymbol{\Sigma}^A$ is a sensible choice and that, in principle, all $\boldsymbol{x}_i$ for $i = 1, \ldots, n$ must be considered in the construction of the prior covariance matrix. Nevertheless, $\boldsymbol{\Sigma}^A$ is a bad choice when censored times are very different and Lemma 3 makes this aspect visible in the extreme case where censored times are of arbitrarily small amount of information in which case $\boldsymbol{\Sigma}^U$ is preferable (although in a context of minimal size, this lemma can be seen as a formalization of the convenience of using $\boldsymbol{\Sigma}^U$ when a subset of the data barely contains information that was illustrated in the motivating example of Section 3).

These results indicate out that, while being unfeasible to exactly reproduce both matrices with a single proposal, a prior matrix that has the ability to mimic the optimal choice in the situations considered above is ideal. This is the type of behavior our proposed matrix $\boldsymbol{\Sigma}^M$ has and that we have illustrated in the previous section. We interpret Lemmas 2 and 3 as giving support to $\boldsymbol{\Sigma}^M$ and that the ensuing prior distribution is endorsed by PM arguments.

Of course, this ability of $\boldsymbol{\Sigma}^M$ is not free and n_u must be at least $k+2$ to ensure the propriety of the prior that uses $\boldsymbol{\Sigma}^M$ – given the experience in the linear model, one should not expect any PM result at this level of minimal sample size –. Hence none of the previous lemmas about PM apply in this case. While not optimal, this situation is not new within the development of PM and, for instance, as shown in Bayarri et al. (2012) for the linear model, none of the priors with covariance matrix (4.6) exist at many PM levels (e.g. if $n \leq k$), simply because this inverse matrix does not exist.

6 Variable selection

Recall that the model $\mathcal{M}_1$ in (2.2) contains all potential k covariates while $\mathcal{M}_0$ only contains the intercept. The variable selection exercise considers these two models and the rest of $2^k - 2$ models corresponding to all possible combinations of which covariates may be relevant for the studied response $\boldsymbol{y}$ and $\boldsymbol{\delta}$.

This variable selection problem with all these entertained models can be compactly formulated using a k binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)$ stating which of the covariates are included: those corresponding with $\gamma_j = 1$, for $j = 1, \ldots, k$. This way, each competing model is denoted $\mathcal{M}_\gamma$ with (centered) design matrix $\widetilde{\boldsymbol{X}}_\gamma$, which is the corresponding $n \times k_\gamma$ ($k_\gamma = \mathbf{1}^T\boldsymbol{\gamma}$) submatrix of the full matrix $\widetilde{\boldsymbol{X}}$ of dimension $n \times k$. Slightly abusing notation $\mathcal{M}_{(1,\ldots,1)} = \mathcal{M}_1$ and $\mathcal{M}_{(0,\ldots,0)} = \mathcal{M}_0$ and $k_{(1,\ldots,1)} = k$.

Let $\pi_\gamma(\beta_0, \sigma, \boldsymbol{\beta}_\gamma)$ be the prior distribution for $\mathcal{M}_\gamma$. Then, the predictive density of $(\boldsymbol{y}, \boldsymbol{\delta})$, under $\mathcal{M}_\gamma$, is

$$m_\gamma(\boldsymbol{y}, \boldsymbol{\delta}) = \int f_\gamma(\boldsymbol{y}, \boldsymbol{\delta} \mid \beta_0, \sigma, \boldsymbol{\beta}_\gamma)\pi_\gamma(\beta_0, \sigma, \boldsymbol{\beta}_\gamma)d\beta_0\, d\sigma\, d\boldsymbol{\beta}_\gamma.$$

The prior for the parameters within each model $\mathcal{M}_\gamma$ is

$$\pi_\gamma(\boldsymbol{\beta}_\gamma, \beta_0, \sigma) = \pi_\gamma(\boldsymbol{\beta}_\gamma \mid \beta_0, \sigma)\pi_\gamma(\beta_0, \sigma).$$

For $\pi_\gamma(\boldsymbol{\beta}_\gamma \mid \beta_0, \sigma)$ we use the prior (4.3) where our proposal for $\boldsymbol{\Sigma}_\gamma$ is $\boldsymbol{\Sigma}^M$ defined in (4.11) replacing $\widetilde{\boldsymbol{X}}$ with $\widetilde{\boldsymbol{X}}_\gamma$. For the prior for common parameters we use $\pi_0(\beta_0, \sigma) = \pi_\gamma(\beta_0, \sigma) = \sigma^{-1}$.

With all the above, the Bayes factor of $\mathcal{M}_\gamma$ against $\mathcal{M}_0$ is given by

$$B_\gamma(\boldsymbol{y}, \boldsymbol{\delta}) = \frac{m_\gamma(\boldsymbol{y}, \boldsymbol{\delta})}{m_0(\boldsymbol{y}, \boldsymbol{\delta})},$$

and the posterior probability of $\mathcal{M}_\gamma$ is

$$p(\mathcal{M}_\gamma \mid \boldsymbol{y}, \boldsymbol{\delta}) = \frac{B_\gamma p(\mathcal{M}_\gamma)}{\sum_{\gamma'} B_{\gamma'} p(\mathcal{M}_{\gamma'})} = \left(1 + \sum_{\gamma' \neq \gamma} \frac{p(\mathcal{M}_{\gamma'})}{p(\mathcal{M}_\gamma)} \frac{B_{\gamma'}}{B_\gamma}\right)^{-1}, \tag{6.1}$$

where $p(\mathcal{M}_\gamma)$ is the prior probability over the model space.

Objective choices for this distribution are the uniform, $p(\mathcal{M}_\gamma) = 1/2^k$, or the hierarchical uniform prior discussed by Scott and Berger (2010): $p(\mathcal{M}_\gamma) \propto 1/\binom{k}{k_\gamma}$. We recommend this last prior, as it accounts for the multiplicity of comparisons as it has been nicely argued by Scott and Berger (2010). Nevertheless, for specific scenarios other choices for $p(\mathcal{M}_\gamma)$ may be more appealing (trying e.g. to force sparsity, Castillo et al., 2015).

7 Breast cancer survival in Castellón (Spain)

Breast cancer is the leading cause of cancer mortality among women. In 1998 the Breast Cancer Registry of Castellón (a province of Spain located on the east coast of the country with approximately half a million inhabitants) was created to asses the importance of this disease, obtaining health indicators (like incidence and survival) that are crucial for health care authorities.

From this registry, we analyze data of women diagnosed with breast cancer during the decade 2004–2013 having a total of $n = 2116$. The closing date of study is the 31st of December, 2015, meaning that women that were alive at that date are treated as censored. With these conditions we observe $n_u = 360$ uncensored observations, implying 82.9% censored ones. The dependent variable is time to event in logarithmic scale.

Following the suggestions of previous institutional reports (Torrella et al., 2005), as potential explanatory variables we have considered two numerical covariates: number of nodes affected (with mode in 0, median 0, mean 2 and maximum 84) and the age at diagnosis (range (25, 97), with mean 58.54 and median 58). In addition, four prognosis binary factors are included in our study: the presence of local recurrences (3.2%), the presence of metastasis (11.3%), estrogenic hormonal receptors (ER being either 0 "negative" or 1 "positive", 79.4% for this last one) and progesterone hormonal receptors (PGR) with the same coding (66.1% for "positive"). These make a total of 6 possible explanatory variables with $2^6 = 64$ competing models.

For each model we have calculated the Bayes Factors defined in (4.2) with our proposed prior (result were quite robust to the choice of the mixing density). For the prior probabilities over the model space we have used the Scott-Berger prior.

In Table 3 we have collected the posterior probabilities of the first two most probable models. The most probable model is the one containing all six variables, closely followed by the model that removes PGR. These two models accumulate a substantial posterior mass (0.94) clearly indicating that there is strong evidence that the variables nodes, age, metasta, recurrence and ER are explanatory variables, while the predictive power

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

Table 3: 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 4: Breast cancer dataset: inclusion probabilities.

of PGR is uncertain. These conclusions are supported as well by the posterior inclusion probabilities – a popular summary in variable selection exercises – defined as the sum of posterior probabilities of models that contain a variable. For our dataset these have been collected in Table 4 where we can clearly see that all variables have very high marginal probabilities, except PGR for which the prior inclusion probability (0.5) has barely changed.

Interestingly, the role of PGR can be further examined using joint (as opposed to the marginal inclusion probabilities above) measures of influence. In particular, one such measure is the probability

$$Pr(PGR \mid \overline{ER}, \boldsymbol{y}, \boldsymbol{\delta})$$

that is, the probability that PGR explains the response if ER is not contained in the true model (this is obtained similarly to the marginal probabilities). In our problem $Pr(PGR \mid \overline{ER}, \boldsymbol{y}, \boldsymbol{\delta}) = 0.94$ indicating that the explanatory power of PGR is absorbed by ER and if this were absent the probability of PGR being a prognostic factor would be very large.

In order to estimate the magnitude of the influence of each variable in (log) time to survival we must examine the model averaged posterior distribution of the regression coefficients. We have represented these distributions in Figure 2 using a histogram-like representation similarly to what it is done in the illustrative example (see the supplementary material). Recall that, except for PGR, there is strong evidence (essentially accumulated in very few models) that these variables are explanatory variables. As a consequence, these histograms are unimodal and are easy to interpret. As expected, the variables nodes, metasta and recurrence have an important negative effect on time to death. Particularly, the presence of metastasis has an estimated effect of -1.76 (posterior mean), implying that, on average, a woman without metastasis multiplies by a factor of 5.6 her expected time of life over a woman with metastasis. On the other hand, a positive ER acts protectively (posterior mean equal to 0.32). Finally, for the reasons explained earlier, PGR has two modes and must be summarized with caution (e.g. means are useless). If this variable is assumed to have an effect, its sign is positive but with as small an impact as PGR (mean of strictly positive values is 0.16). In Table 5 we have summarized probabilities of survival calculated with the posterior predictions with certain values of the covariates.

It is also interesting to analyze the estimation of the effective sample $N_{(\beta_0,\sigma)}$. The histogram corresponding to its posterior distribution averaged over the competing mod-

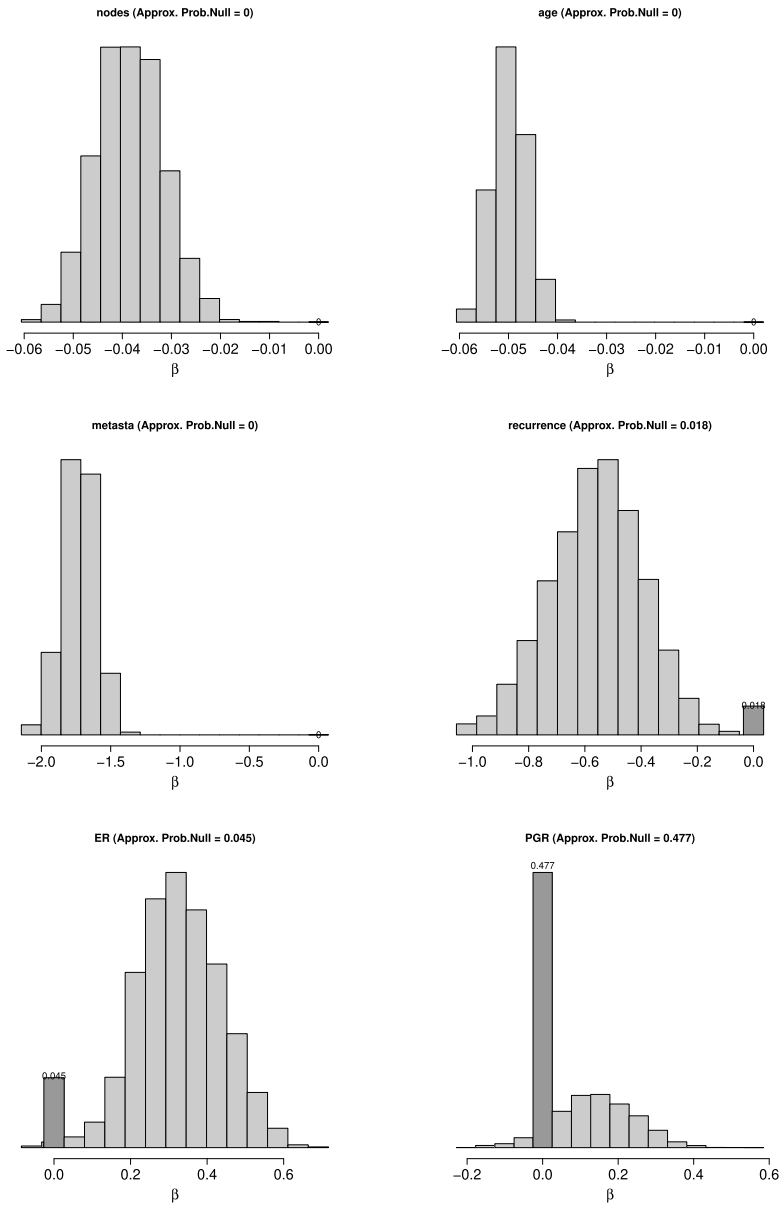

Figure 2: Breast cancer dataset. Model averaged posterior distributions of the regression coefficients for each potential covariate. Dark gray area represents the probability of no effect and the light gray area the distribution of probability given there is an effect.

recurrence	metasta	nodes	age	ER	PGR	Survival at year 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 5: Model averaging estimation of survival probabilities for different values of explanatory variables. Last row is for an average case (values of the covariates at the sample mean).

els is represented in Figure 3. The posterior mean of this parameter is 714 uncensored individuals with a standard deviation of 32. This can be interpreted as one uncensored individual having similar information content to roughly five censored datum.

8 A simulation study over heart transplant data

In order to deep in the comparison, we have designed a simulated experiment obtained from the heart transplant data set analyzed in the supplementary material. In this data set $n = 69$ and in our experiment, for each person in the study, date of death is simulated as the original date entering the study plus the simulation of a log-Normal distribution whose logarithm has the following mean and standard deviation:

$$mean_i = 8.35 - 0.04 \times \text{age}_i + 0 \times \text{spur1}_i + 0 \times \text{spur2}_i, \quad sd_i = 0.6.$$

The parameters are set to the estimated values in the real data set to keep the experiment close, as much as possible, to the real experiment. The variables spur1 and spur2 act as spurious variables (do not have any effect on the response) and are simulated as independent normal standard random variables. The closing date for the study is the same as in the original data set: April 1, 1974.

With the above scheme we simulated 50 data sets to which we performed variable selection based on BF_{robust}, TBF_{EB} and TBF_{ZS}. Inclusion probabilities in the form of cloud of points are represented in Figure 2 (supplementary) and further summarized in Figure 3 (supplementary). The method based on TBF_{EB} is clearly much less conservative (larger proportion of true positives but also false positives) than the one based on TBF_{ZS} and BF_{robust}, which remarkably, produce quite similar results (TBF_{ZS} being, almost systematically, slightly more conservative than BF_{robust}; the differences enlarge when the inclusion probabilities are close to 1/2). This numerical study is far from being exhaustive and more work (particularly theoretically) is needed to understand the

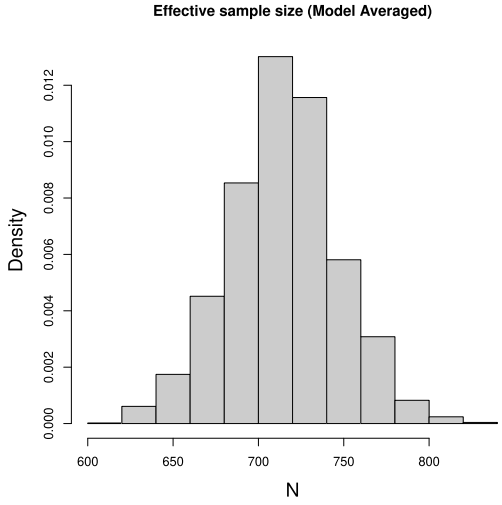

Figure 3: Breast cancer dataset. Model averaged posterior distributions of the effective sample size $N_{(\beta_0,\sigma)}$.

connections between actual Bayes factors and sensible approximations as those based on the TBFs in problems with incomplete information (censored observations in this case). In our opinion, the close agreement between BF_{robust} and TBF_{ZS} that we have seen in the experiment, reinforces the main messages motivating both approaches. On the one side it empirically confirms the Bayesian nature of TBF_{ZS} (our prior could be interpreted as the explicit prior behind it) and on the other side, highlights the adequacy of the route that we have followed in this paper to handle, from a strict Bayesian point of view, a variable selection problem with censored data.

9 Further remarks

In this paper we have developed a full Bayesian approach to the problem of variable selection based on model selection when the observations are subject to censoring. In this context, we have emphasized the importance of using adequate generalizations of standard objective variable selection priors that, in survival problems, take into account the different information content in observational units. We have seen, in a motivating example and from a predictive matching perspective, the importance of such generalizations particularly in situations where the distribution of covariates substantially differs in censored and uncensored observations. A referee has pointed out a close correspondence between this situation and the mechanism of *missing at random* in missing data where covariates and missingness (here censoring) are related. Following this parallelism, in a *missing completely at random* context (hence the propensity of being censored is not related with the explanatory variables) standard priors for variable selection (e.g. g-priors with $\mathbf{\Sigma}^A$) are expected to produce sensible results. Of course, it is difficult to

Variable	nodes	age	metasta	recurrence	ER	PGR
BF_{robust}	1.00	1.00	1.00	0.98	0.96	0.52
TBF_{EB}	1.00	1.00	1.00	0.99	0.96	0.54
TBF_{ZS}	1.00	1.00	1.00	0.98	0.96	0.49

Table 6: Breast cancer dataset: inclusion probabilities with our proposed approach and those obtained with the Test-Based Bayes factors.

anticipate in a given problem whether we are in one or other situation but the approach we propose has the ability (partially because our setting works conditionally on known censoring times, c_i, and covariates) of being automatically self-adaptive without the need to specifically model the mechanisms that underlie censoring.

Our proposal develops in the context of g priors, and we have proposed the matrix $\mathbf{\Sigma}^M$ in (4.11) as a convenient prior covariance matrix arguing that it automatically adjusts for this varying information among units. Throughout the paper, we have largely discussed about the benefits of using this matrix as opposed to other naive implementations of the g prior that either use equally all units ($\mathbf{\Sigma}^A$) or only the uncensored ones ($\mathbf{\Sigma}^U$).

We now compare our proposal with the method for variable selection using test-based Bayes factors (TBFs) originally proposed by Johnson (2008); Hu and Johnson (2009) later revisited by Held et al. (2015), and recently developed in the Cox model by Held et al. (2016). Within this approach, the Bayes factors B_γ in (6.1) are replaced by the ratio of the integrated likelihoods for the deviance statistic under M_γ and under M_0 which, appealingly, result in simple functions of the deviance. Despite its strong Bayesian origin, the resulting method is not a full Bayesian procedure both because the real model is not used (but the distribution of deviance statistic under M_γ) but also because the implicit nature of the prior for the regression parameters makes it difficult, in many cases, to asses its strict validity as a prior distribution (i.e. the implicit prior covariance for the regression parameters is the expected information matrix (4.8) which itself depends on $\boldsymbol{\beta}$).

TBFs depend on a hyper-parameter g' (equivalent to our $g\,n$) that can be specified in several ways. We here consider the specifications that seem to be closer competitors to our approach, namely g' estimated via Empirical Bayes (labeled TBF_{EB}) and the adapted Zellner-Siow prior based on the number of uncensored observations n_u (labeled TBF_{ZS}). We refer the reader to the references provided for further details.

For the Breast cancer data set we have collected in Table 6 the inclusion probabilities based on TBF_{EB} and TBF_{ZS} that we compare with our results (based on the mixing prior (4.5) that we label BF_{robust}). The results are quite similar, BF_{robust} being between both TBFs.

Our full Bayesian methodology, admittedly, is computationally very demanding so, as it is, it's only suitable for small to moderate k. In this regard, a main bottleneck in our approach is a covariance matrix $\mathbf{\Sigma}^M$ that is unknown since it depends on (β_0, σ) making it quite more challenging computing the integrals defining the Bayes factors. To

have an idea of the computational burden of our approach, running the real example in Section 7 took approximately 40 hours in a Linux machine with 32 parallel treads.

An approximate implementation of our methodology with an Empirical Bayes flavor is using (for each entertained model) $\mathbf{\Sigma}_\gamma^M(\hat{\beta}_0, \hat{\sigma})$ as the prior covariance matrix, where $(\hat{\beta}_0, \hat{\sigma})$ are the posterior means of the corresponding parameters under the null model. In order to speed up the computation, this strategy can be accompanied with the evaluation of the integral in (4.1) using Laplace (several authors have argued about the convenience of Laplace approximations over numerical approaches like for instance, Sabanes and Held, 2011). The only parameter that cannot properly handled via Laplace integration is g (and particularly with the robust prior) since, depending on the model, it may have a mode in the boundaries of its parametric space making the Laplace approximation very poor. Hence, in the case of random g, this parameter is integrated with standard numerical quadrature.

We have implemented the above described numeric strategy in R and the code is available in a public github repository.[3] The inclusion probabilities for the breast cancer dataset in Section 7 turned out to be very similar to the exact values (values 1, 1, 1, 0.98, 0.95 and 0.49 in the order that appears in Table 6) but taking only 3 minutes to compute them (a reduction of 99.8% of computational time). We also have computed this approximation for the 50 simulated data sets constructed upon the heart data set as is described above. Computational time was reduced from 850 minutes using importance sampling to 5 minutes (94% of reduction in the computational time) using this Laplace approximation. The results showed the accuracy of the approximation since the maximum difference in absolute value we observed between the 150 inclusion probabilities was 0.06.

This approach is respectful with the overall message in this paper and the resulting posterior distribution over the model space could be easily explored with Gibbs sampling schemes (see e.g. Garcia-Donato and Martinez-Beneito, 2013) making it feasible to face problems with larger k that we'll consider in future research.

Supplementary Material

Supplementary material for: "A Model Selection Approach for Variable Selection with Censored Data" (DOI: 10.1214/20-BA1207SUPP; .pdf). Proofs of theorems and lemmas. Details about BF calculation and on simulation study.

References

Antoniadis, A., Fryzlewicz, P., and Letué, F. (2010). "The Dantzig Selector in Cox's Proportional Hazards Model." *Scandinavian Journal of Statistics*, 37(4): 531–552. MR2779635. doi: https://doi.org/10.1111/j.1467-9469.2009.00685.x. 273

[3] https://github.com/scabras/censBayesVarsel.

Barbieri, M. M. and Berger, J. O. (2004). "Optimal Predictive Model Selection." *The Annals of Statistics*, 32(3): 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 276

Bayarri, M., Berger, J., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian Model Choice with Application to Variable Selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 272, 278, 279, 280, 285, 287, 289

Bayarri, M. and García-Donato, G. (2008). "Generalization of Jeffreys Divergence-Based Priors for Bayesian Hypothesis Testing." *Journal of the Royal Statistical Society: Series B*, 70(5): 981–1003. MR2530326. doi: https://doi.org/10.1111/j.1467-9868.2008.00667.x. 281

Bayarri, M. J. and García-Donato, G. (2007). "Extending Conventional Priors for Testing General Hypotheses in Linear Models." *Biometrika*, 94(1): 135–152. MR2367828. doi: https://doi.org/10.1093/biomet/asm014. 273

Berger, J. (2006). "The Case for Objective Bayesian Analysis." *Bayesian Analysis*, 1(3): 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 272

Berger, J., Bayarri, M., and Pericchi, L. (2014). "The effective sample size." *Econometric Reviews*, 33(1-4): 197–217. MR3170846. doi: https://doi.org/10.1080/07474938.2013.807157. 281

Berger, J. and Pericchi, L. (2001). "Objective Bayesian Methods for Model Selection: Introduction and Comparison." In Lahiri, P. (ed.), *Model Selection*, volume 38, pp. 135–207. Institute of Mathematical Statistics. URL http://www.jstor.org/stable/4356165. MR2000753. doi: https://doi.org/10.1214/lnms/1215540968. 272, 273, 287

Berger, J. and Pericchi, L. (2004). "Training samples in objective Bayesian model selection." *The Annals of Statistics*, 32: 841–869. MR2065191. doi: https://doi.org/10.1214/009053604000000238. 273

Berger, J. O. and Pericchi, L. R. (1996). "The Intrinsic Bayes Factor for Model Selection and Prediction." *Journal of the American Statistical Association*, 91(433): pp. 109–122. URL http://www.jstor.org/stable/2291387. MR1394065. doi: https://doi.org/10.2307/2291387. 273

Cabras, S., Castellanos, M., and Perra, S. (2014). "Comparison of objective Bayes factors for variable selection in parametric regression models for survival analysis." *Statistics in Medicine*, 33(26): 4637–4654. MR3267387. doi: https://doi.org/10.1002/sim.6249. 273, 286

Cabras, S., Castellanos, M., and Perra, S. (2015). "A new minimal training sample scheme for intrinsic Bayes factors in censored data." *Computational Statistics & Data Analysis*, 81: 52–63. MR3257400. doi: https://doi.org/10.1016/j.csda.2014.07.012. 273

Candes, E. and Tau, T. (2007). "The Dantzig Selector: Statistical Estimation When

p Is Much Larger than n." *The Annals of Statistics*, 35(6): 2313–2351. MR2382644. doi: https://doi.org/10.1214/009053606000001523. 273

Castellanos, M. E., García-Donato, G. and Cabras, S. (2020). "A Model Selection Approach for Variable Selection with Censored Data – Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/20-BA1207SUPP. 274

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 290

Clyde, M. (1999). "Bayesian Model Averaging and Model Search Strategies." In Bernardo, J., Berger, J., Dawid, A., and Smith, A. (eds.), *Bayesian Statistics*, volume 6, 157–185. Oxford University Press. MR1723497. 281

Consonni, G., Fouskakis, D., Liseo, B., and Ntzoufras, I. (2018). "Prior distributions for objective Bayesian analysis." *Bayesian Analysis*, 13: 627–679. MR3807861. doi: https://doi.org/10.1214/18-BA1103. 272

De Santis, F., Mortera, J., and Nardi, A. (2001). "Jeffreys priors for survival models with censored data." *Journal of Statistical Planning and Inference*, 99(2): 193–209. MR1865291. doi: https://doi.org/10.1016/S0378-3758(01)00080-5. 272

DiCiccio, T., Kass, R., Raftery, A., and Wasserman, L. (1997). "Computing Bayes factors by combining simulation and asymptotic approximations." *Journal of American Statistical Association*, 92: 903–915. MR1482122. doi: https://doi.org/10.2307/2965554. 287

Fan, J. and Li, R. (2002). "Variable selection for Cox's proportional hazards model and frailty." *Annals of Statistics*, 30(1): 74–99. MR1892656. doi: https://doi.org/10.1214/aos/1015362185. 273

Fernández, C., Ley, E., and Steel, M. (2001). "Benchmark Priors for Bayesian Model Averaging." *Journal of Econometrics*, 100: 381–427. MR1820410. doi: https://doi.org/10.1016/S0304-4076(00)00076-2. 279

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). "Power-Expected-Posterior priors for generalized linear models." *Bayesian Analysis*, 13(3): 721–748. MR3807864. doi: https://doi.org/10.1214/17-BA1066. 273

Garcia-Donato, G. and Forte, A. (2018). "Bayesian Testing, Variable Selection and Model Averaging in Linear Models using R with BayesVarSel." *The R Journal*, 10(1): 155–174. 276

Garcia-Donato, G. and Martinez-Beneito, M. (2013). "On Sampling strategies in Bayesian variable selection problems with large model spaces." *Journal of the American Statistical Association*, 108(501): 340–352. MR3174624. doi: https://doi.org/10.1080/01621459.2012.742443. 296

George, E. and McCulloch, R. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 272

Held, L., Gravestocka, I., and Sabanés Bové, D. (2016). "Objective Bayesian model selection for Cox regression." *Statistics in Medicine*, 35: 5376–5390. MR3573064. doi: https://doi.org/10.1002/sim.7089. 272, 295

Held, L., Sabanés Bové, D., and Gravestock, I. (2015). "Approximate Bayesian model selection with the deviance statistic." *Statistical Science*, 30: 242–257. MR3353106. doi: https://doi.org/10.1214/14-STS510. 295

Hu, J. and Johnson, V. (2009). "Bayesian model selection using test statistics." *Journal of the Royal Statistical Society. Series B (Methodological)*, 71(1): 143–158. MR2655527. doi: https://doi.org/10.1111/j.1467-9868.2008.00678.x. 295

Jeffreys, H. (1961). *Theory of Probability*. Oxford University Press, 3rd edition. MR0187257. 278, 279, 287

Johnson, V. (2008). "Properties of Bayes factors based on test statistics." *Scandinavian Journal of Statistics*, 35(2): 354–368. MR2418746. doi: https://doi.org/10.1111/j.1467-9469.2007.00576.x. 272, 295

Johnson, V. and Rossell, D. (2010). "On the use of non-local prior densities in Bayesian hypothesis tests." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2): 143–170. URL http://dx.doi.org/10.1111/j.1467-9868.2009.00730.x. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 272, 277

Kass, R. and Raftery, A. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 271, 281

Kass, R. and Wasserman, L. (1995). "A Reference Bayesian Test for Nested Hypotheses and its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association*, 90(431): 928–934. MR1354008. 279

Kass, R. and Wasserman, L. (1996). "The selection of prior distributions by formal rules." *Journal of the American Statistical Association*, 91(435): 1343–1369. 272

Li, Y. and Clyde, M. (2018). "Mixtures of g-priors in generalized linear models." *Journal of the American Statistical Association*, 113: 1828–1845. MR3902249. doi: https://doi.org/10.1080/01621459.2018.1469992. 273

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). "Mixtures of g-Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 272

Maruyama, Y. and Strawderman, W. (2010). "Robust Bayesian variable selection with sub-harmonic priors." Technical report, arXiv:1009.1926. 273

Moreno, E., Bertolino, F., and Racugno, W. (1998). "An intrinsic limiting procedure for model selection and hypothesis testing." *Journal of the American Statistical Association*, 93: 1451–1460. MR1666640. doi: https://doi.org/10.2307/2670059. 273

Nikooienejad, A. and Johnson, V. (2018). *BVSNLP: Bayesian variable selection in high dimensional settings using nonlocal priors*. 276

Nikooienejad, A., Wang, W., and Johnson, V. (2018). "Bayesian Variable Selection For Survival Data Using Inverse Moment Priors." Technical Report arXiv:1712.02964, Cornell University. 272, 277

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): pp. 99–138. URL http://www.jstor.org/stable/2346088. MR1325379. 273

Perra, S., Cabras, S., and Castellanos, M. (2013). *Objective Bayesian Variable Selection for Censored Data*. LAP Lambert Academic Publishing. 273

Sabanes, D. and Held, L. (2011). "Hyper-g priors for generalized linear models." *Bayesian Analysis*, 6: 387–410. MR2843537. doi: https://doi.org/10.1214/ba/1339616469. 273, 296

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. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 290

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. 272

Steel, M. (2019). "Model Averaging and its use in economics." *Journal of Economic Literature (forthcoming)*. 274

Torrella, A., Martínez-Beneito, M., Alacreu, M., M., S., and Guallar, E. (2005). "Informes de Salud 85. Incidencia y supervivencia del cáncer de mama femenino en la proovincia de Castellón 1995–2002." Generalitat Valenciana. 290

Touloupou, P., Alzahrani, N., Neal, P., S.E.F., S., and T. J., M. (2018). "Efficient model comparison techniques for models requiring large scale data augmentation." *Bayesian Analysis*, 13(2): 437–459. MR3780430. doi: https://doi.org/10.1214/17-BA1057. 286

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions." In Zellner, A. (ed.), *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, 389–399. Edward Elgar Publishing Limited. MR0881437. 272, 279

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. (eds.), *Bayesian Statistics 1*, 585–603. Valencia: University Press. 272, 279

Zhang, H. and Lu, W. (2007). "Adaptive Lasso for Cox's Proportional Hazards Model." *Biometrika*, 94(3): 691–703. URL http://www.jstor.org/stable/20441405. MR2410017. doi: https://doi.org/10.1093/biomet/asm037. 273

Acknowledgments

The authors would like to thank the Epidemiological Unit and the cancer registry of Castellón of the Conselleria de Sanitat Universal i Salut Pública for sharing the breast cancer dataset.

The authors also thank the Editor, an Associate Editor and a referee for very valuable suggestions that have greatly improved the manuscript.