Received: 24 January 2022 | Accepted: 25 November 2022

DOI: 10.1111/biom.13823

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Model uncertainty quantification in Cox regression

Gonzalo García-Donato[1] | Stefano Cabras[2] | María Eugenia Castellanos[3]

[1]Department of Economy and Finance, University of Castilla-La Mancha, Albacete, Spain

[2]Department of Statistics, Carlos III University of Madrid, Getafe, Madrid, Spain

[3]Department of Informatics and Statistics, Rey Juan Carlos University, Móstoles, Madrid, Spain

Correspondence
María Eugenia Castellanos, Department of Informatics and Statistics, Rey Juan Carlos University, Móstoles, Madrid, Spain.
Email: maria.castellanos@urjc.es

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: Grant PID2019-104790GB-I00 funded by MCIN/AEI

Abstract

We consider covariate selection and the ensuing model uncertainty aspects in the context of Cox regression. The perspective we take is probabilistic, and we handle it within a Bayesian framework. One of the critical elements in variable/model selection is choosing a suitable prior for model parameters. Here, we derive the so-called conventional prior approach and propose a comprehensive implementation that results in an automatic procedure. Our simulation studies and real applications show improvements over existing literature. For the sake of reproducibility but also for its intrinsic interest for practitioners, a web application requiring minimum statistical knowledge implements the proposed approach.

KEYWORDS

Bayesian variable selection, conventional prior, Fisher information, median model, survival analysis

1 | INTRODUCTION

In survival analysis, inferences and predictions are sensitive to the form assumed for the risk function (the instantaneous possibility of the event). One possible strategy for overcoming this issue is to treat the actual risk function as unknown, resulting in more robustness to misspecifications. One such method is the Cox regression, a semi-parametric approach widely used in many applied disciplines such as epidemiology, and which will be the focus of our work.

The Cox model assumes that the explanatory variables are known. When this hypothesis is relaxed, we obtain a more accurate representation of reality, called model uncertainty, which we consider in this paper. The resulting procedures are closely connected to variable selection and model averaging methods.

As an illustrative approximation to model uncertainty, which we consider throughout the paper, we use the primary biliary cholangitis (PBC) dataset, previously analyzed by several authors such as Hoeting et al. (1999a) and more recently by Fleming and Harrington (2011). In this study, time to death or liver transplant, Y, is observed in a set of patients subject to possible censoring (e.g., people who withdraw from the experiment), together with a bunch of $p = 15$ variables including age and several other potential prognostic indicators (described in Appendix F). This survival study initially enrolled $n = 312$ patients, of which $n_u = 125$ were uncensored (59% censoring). The main interest is to obtain an estimation of the survival function for different patient profiles. This estimation is a simple statistical exercise once the covariates enter the model are specified. Ideally, we would like to use only the true explanatory variables, but unfortunately, we do not know which ones are. We could hinder this choice

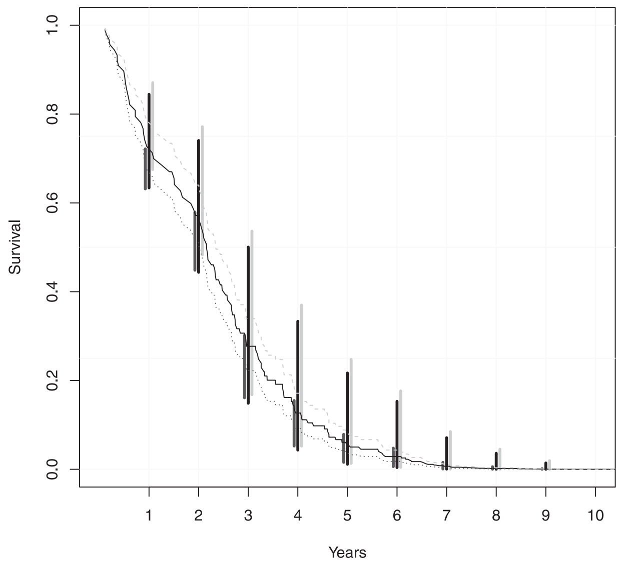

FIGURE 1 For the primary biliary cholangitis (PBC) dataset, survival functions and 95% credible intervals for the different years based on the model with all covariates (dashed line; light gray interval); on the highest posterior probability model (points; dark gray interval) and all models weighted by their posterior probability (solid line; black interval). The result corresponds to an imaginary patient with a bad prognosis (covariates age; bili; edema and protime set at the 90% percentile; covariate albumin set at 10% and the rest fixed at the median value).

and proceed as if all variables were truly affecting the evolution of the disease, obtaining the dashed curve represented in Figure 1 with 95% credible intervals (obtained for the consecutive years) colored in light gray. An alternative, two-step strategy would be to use a variable selection method and then proceed with the estimation based on these variables. In the PBC dataset, the model supported with the highest posterior probability (to be defined later) contains only four variables. Using these, we have represented the survival function in the same figure, as the curve made of points and the intervals colored in dark gray. The differences between the two estimated functions are visible. The substantially lower precision when all the variables are used is due to many more regression parameters participating in the survival function estimation. The result corresponding to the two-step procedure could be more accurate, given that the model used is expected to be "close to the truth". Unfortunately, the truth is unknown, and problems with even moderate p are subject to considerable model uncertainty that rarely dissipates with data. Every single model, including the most promising ones, receives a little support from the data and inferences obtained based on a single model are questionable. The PBC example illustrates this situation. We will see that the best models have small probabilities, demonstrating the need for methodologies that account for the uncertainty concerning the selected model.

An approach based on model uncertainty would average the estimated survival curve provided by each possible model (each combination of potential covariates), weighted by a certain measure of evidence in favor of the possible models.

The obvious choice for such weighting measures within the Bayesian paradigm is the posterior probabilities of models, the primary tool in Bayesian model uncertainty methods. In Figure 1, the resulting model uncertainty-based survival curve has been plotted as a solid line with credible intervals colored in black for the illustrative example. We observe that point estimates compromise the two procedures discussed and exhibit variability similar to that obtained using all the variables. This result is a direct consequence of the considerable uncertainty present in this dataset (shown in our findings, but also suggested in previous studies; see, e.g., Hoeting et al., 1999a).

The Bayesian construction of posterior probabilities of models is based on Bayes factors and has its origins in the significance tests introduced by Jeffreys (1961). The interested reader is referred to Etz and Wagenmakers (2017) for a historical perspective on the development of Bayes factors and Robert et al. (2009) for a modern review of Jeffreys' book. The resulting posterior probabilities can be used formally within a purely probabilistic view or more casually as appropriate weights to draw more realistic inferences that account for model uncertainty.

Posterior probabilities are supported by many appealing properties summarized in Berger and Pericchi (2001). Notably, the Bayesian approach is automatic Ockham's razor—hence favoring the simpler models for a similar fit—and provides straightforward solutions to control for multiplicity. This problem arises when considering an extensive list of models.

The general guidelines for applying the Bayesian approach to model uncertainty are established. These are briefly introduced in Section 2. Nevertheless, *the devil is in the details* and the implementation in particular problems is plenty of difficulties that need meticulous considerations. In this paper, all these challenges are tackled in the context of Cox regression models, and the result is a fully Bayesian procedure that is automatic (free of tuning parameters). This procedure behaves very competitively compared with existing procedures, as shown in an extensive simulation study in Appendix D in the Supplementary information. For reproducibility and accessibility, we provide accompanying software implemented in R functions and a shiny application.

The main theoretical contribution in the paper is a novel prior distribution specifically derived for Cox regression models. Its development, detailed in Section 3, follows the tradition of the so-called conventional (or g-priors) introduced by Zellner and Siow (1980). These priors are model-specific because they use the expected information matrix of the models to form the prior variance. By construction, g-priors induce dependence among regression parameters, a key feature as recently shown by Barbieri et al. (2021). An essential part of this research is the derivation of this matrix in Cox regression models and the study of its properties in the context of model choice. Surprisingly, this matrix has a manageable expression and an appealing interpretation as a weighted covariance matrix. Once the priors are defined, the questions related to the numerical implementation emerge. In Section 4, we provide instructions to approximate the integrals defining the Bayes factors with Laplace numerical quadrature. This technique solves the numerical question within each model, but the most challenging feature of model uncertainty problems comes from the enormous cardinality, 2^p, of possible models. This problem has been called in the literature "model search", and we approach it using a simple Gibbs sampling scheme. This algorithm was originally mentioned in the context of linear regression by George and McCulloch (1997) and its properties are studied in depth in Garcia-Donato and Martinez-Beneito (2013). In contrast to the predominant procedures aimed at searching for the best models, the proposed one samples models according to (approximately) their posterior probability, preserving the problem's probabilistic structure. While the benefits of sampling methods in selecting a single model are perhaps debatable, their advantages in handling model uncertainty problems are evident since they automatically propagate the variability following the standards of probability laws. The procedure has been called Bayesian model averaging, which we illustrate in the context of the survival study underlying the PBC dataset in Section 5.

The problem we consider has recently been studied by Nikooienejad et al. (2020) from a similar perspective. However, the two procedures substantially differ in fundamental aspects concerning the implementation of the Bayesian method. First, the prior in Nikooienejad et al. (2020) is a product of the non-local priors introduced by Johnson and Rossell (2010, 2012) without any connection with the model at hand (in this case, the Cox model) and its apparent particularities. Second, their proposal for "model search" is based on optimization methods, specifically conceived to discover the most promising models. Because this is not a sampling procedure, the method must rely on re-normalization to provide the collected models by some probabilistic structure and produce estimations with a model uncertainty flavor. Garcia-Donato and Martinez-Beneito (2013) documented that re-normalization-based methods produce unsatisfactory estimations in model uncertainty procedures. These two main differences may well explain why our proposal generally outperforms that in Nikooienejad et al. (2020) as we show in the numerical study in Appendix D in the Supplementary information. In particular, it performs better in estimating unknown parameters and producing fewer false positives. A couple of decades ago, Hoeting et al. (1999a) also proposed model uncertainty methods within the Cox model. Their methods are based on the Bayesian information criterion (BIC) and not on an actual Bayes factor. When compared to Hoeting et al. (1999a), both our method and Nikooienejad et al. (2020) outperformed it. As we will show, the BIC-based procedure produces more false positives and true positives but with a considerable error in estimating the model parameters.

2 | BASIC STATISTICAL METHODS

This part is devoted to the essential ingredients used throughout the paper. Section 2.1 introduces the Cox regression model when the set of explanatory variables is known. Section 2.2 presents the general aspects of the problem of model uncertainty caused by the uncertainty about which covariates are influential.

2.1 | Cox semiparametric model

Suppose we have p explanatory variables $\{x_1, x_2, \dots, x_p\}$ known to be relevant in explaining the time-to-terminal-event response variable $Y \in \mathbb{R}^+$. This outcome is subject to right censoring, meaning that if Y exceeds a known censoring time $C \in \mathbb{R}^+$, its real value is unobserved, leading to a *right* censored observation. In survival analysis, Y could be the time to death, illness relapse, disease progression, or any other study or clinical trial endpoint. In practice, these survival times are censored because, in studies or clinical trials, all patients are not followed or may abandon the study.

We use the standard notation in survival analysis by denoting with y_i the observed time-to-event for individual $i = 1, \dots, n$, which is only observed if $y_i < c_i$, where c_i is the censoring time. We denote δ_i as the observed binary variable that records a one, $\delta_i = 1$ if $y_i < c_i$ (uncensored) and $\delta_i = 0$ otherwise (censored). Once the experiment has finished, we observe which individuals have or have not been censored in the vector $\boldsymbol{\delta}^\top = (\delta_1, \dots, \delta_n)$. For those $n_u = \sum_{i=1}^n \delta_i$ uncensored times, we observe their survival times $(y_1, \dots, y_{n_u})$. In this notation, we assume, without loss of generality, that uncensored observations correspond to the first individuals $\{1, \dots, n_u\}$ of a sample of $n > n_u$ individuals. Likewise, we denote the number of censored observations by $n_c = n - n_u$. Although it is arguably less clear, we find it useful, for the sake of simplicity in the formulas, to denote $\boldsymbol{y}^\top = (y_1, \dots, y_{n_u}, c_{n_u+1}, \dots, c_n)$. For each i, the p-dimensional vector $\boldsymbol{x}_i$ contains the observed values of the covariates. Finally, $\boldsymbol{X}$ is the $n \times p$ matrix whose ith row is $\boldsymbol{x}_i^\top$.

The initial hypothesis underlying Cox proportional models assumes a common (to all models) baseline hazard function for all individuals, $h_0(y)$ (i.e., the ratio of the probability density function to the survival function, and the complement of the cumulative distribution function). What makes the individual hazard, h_i, different is the multiplicative effect of the covariates in the form:

$$\begin{aligned} h_i(y_i) &= h_0(y_i)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) \\ &= h_0(y_i)\exp(\mathbf{x}_i^\top \boldsymbol{\beta}). \end{aligned} \tag{1}$$

The cumulative risk function, $H_0(y) = \int_0^y h_0(t)\,dt$, governs the probability of the terminal event being censored since the larger the risk, the lower the probability of being censored. Within the semi-parametric Cox model, h_0 is treated non-parametrically, thus inducing the so-called partial likelihood Cox (1972):

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{\sum_{j\in R(y_i)} \exp(\mathbf{x}_j^\top \boldsymbol{\beta})} \right]^{\delta_i}, \tag{2}$$

where the subset $R(t) = \{i : y_i \geq t\}$ denotes the subjects that are still present in the study at time t—censored or uncensored, that is, subjects that are at risk of experiencing the terminal event at time t. As discussed by Johansen (1983), from a frequentist perspective, $L_p(\boldsymbol{\beta})$ is obtained by maximizing the likelihood as a function of $h_0(t)$ only, with the $\boldsymbol{\beta}$ parameters assumed fixed. From a Bayesian perspective, Equation (2) is the marginal likelihood obtained by marginalizing $h_0(t)$ over a Gamma process (see Kalbfleisch 1978 and also Bailey 1983; Murphy and Vaart 2000; Sinha et al. 2003 for related arguments).

2.2 | Basic formulae in model uncertainty

In the vast majority of applied studies, it is unknown which, if any, of $\{x_1, \ldots, x_p\}$, have a real impact on the response. Therefore, the true model is unknown, leading to a situation subject to model uncertainty. The list of possible models can be expressed by introducing a binary parameter vector, $\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_p)$, where $\gamma_j = 1$ if the response depends on x_j, and zero otherwise. For example, the model that includes only x_3 corresponds to $\boldsymbol{\gamma}^\top = (0, 0, 1, 0, \ldots, 0)$.

Each of the above competing models is labeled as M_γ and the set that contains all of them is the model space $\mathcal{M}$. The cardinality of this set containing only main effects is 2^p and, traditionally, the model that contains all covariates is referred to as the "full model" with associated likelihood (2). In the other extreme, the "null model" (M_0) does not contain any covariates and its likelihood is $L_p(M_0) = \prod_{i=1}^n (\#R(y_i))^{-\delta_i}$ (cf. Equation (2) with $\boldsymbol{\beta} = \mathbf{0}$, here $\#A$ indicates the cardinality of set A).

The rest of the models can be expressed as:

$$L_p(M_\gamma, \boldsymbol{\beta}_\gamma) = \prod_{i=1}^{n} \left[\frac{\exp\left(\mathbf{x}_{i,\gamma}^\top \boldsymbol{\beta}_\gamma\right)}{\sum_{j\in R(y_i)} \exp\left(\mathbf{x}_{j,\gamma}^\top \boldsymbol{\beta}_\gamma\right)} \right]^{\delta_i}, \tag{3}$$

where $\boldsymbol{\beta}_\gamma$ and $\mathbf{x}_{i,\gamma}$ are the components of $\boldsymbol{\beta}$ and $\mathbf{x}_i$ with a one in $\boldsymbol{\gamma}$, so the dimension of both vectors is $k_\gamma = \sum \gamma_i$, the number of regression coefficients under M_γ, also referred to as the *model size*.

The posterior distribution assigns to each model M_γ its probability conditional on the available data:

$$p(M_\gamma \mid \mathbf{y}) = \frac{m_\gamma(\mathbf{y}) p(M_\gamma)}{\sum_{\gamma\in\mathcal{M}} m_\gamma(\mathbf{y}) p(M_\gamma)}, \tag{4}$$

where $p(M_\gamma)$ is the prior probability of M_γ and $m_\gamma(\mathbf{y})$ is the corresponding prior predictive marginal density:

$$m_\gamma(\mathbf{y}) = \int L_p(M_\gamma, \boldsymbol{\beta}_\gamma)\,\pi_\gamma(\boldsymbol{\beta}_\gamma)\,d\boldsymbol{\beta}_\gamma, \tag{5}$$

where $\pi_\gamma(\boldsymbol{\beta}_\gamma)$ are the prior distributions for the model-specific parameters in M_γ.

Alternatively, the posterior distribution can be expressed using Bayes factors, which are the ratio of prior predictive marginals for two different models. Therefore, if we divide the numerator and the denominator in expression (4) by the marginal of a fixed model (e.g., M_0), we obtain:

$$p(M_\gamma \mid \mathbf{y}) = \frac{B_\gamma(\mathbf{y}) p(M_\gamma)}{\sum_\gamma B_\gamma(\mathbf{y}) p(M_\gamma)}, \tag{6}$$

where $B_\gamma(\mathbf{y})$ is the Bayes factor of M_γ to M_0, that is, $B_\gamma(\mathbf{y}) = m_\gamma(\mathbf{y})/m_0(\mathbf{y})$.

These posterior probabilities assign weights, based on the evidence provided by the data, to each of the models and are the crucial tool for tackling model uncertainty problems. There are popular summaries of this distribution such as the highest posterior probability model (HPM), that is, HPM $:= \arg\max_{\gamma\in\mathcal{M}} p(M_\gamma \mid \mathbf{y})$; the posterior inclusion probabilities (which for the jth covariate is $p_j = \sum_{\gamma\in\mathcal{M}:\gamma_j=1} p(M_\gamma \mid \mathbf{y})$) or the median probability model (MPM), which is the model containing the covariates with $p_j > 0.5$ (Barbieri et al., 2021; Barbieri and Berger, 2004). A more interesting usage for model uncertainty purposes is that $p(M_\gamma \mid \mathbf{y})$ can easily be used to define inferences that account for model uncertainty. If Δ is an unknown quantity whose estimation under M_γ

is $\widehat{\Delta}_\gamma$, then:

$$\widehat{\Delta} = \sum_{\gamma \in \mathcal{M}} \widehat{\Delta}_\gamma \, p(M_\gamma \mid \boldsymbol{y}) \tag{7}$$

is an estimation that weights the contribution of each model estimation using its posterior probability. Many authors have called the above approach to inference model averaging (see Hoeting et al., 1999b; Raftery et al., 1997; Steel, 2020, for key references on model averaging). In the above expression, $\widehat{\Delta}_\gamma$ can be a probabilistic distribution (e.g., a posterior or a predictive density) leading to $\widehat{\Delta}$ being a discrete mixture of distributions.

Only quantities with a common meaning across the different models can obtain model uncertainty estimations. In the context of survival analysis, the survival function $S(y) = P(Y > y)$ is one of such parameters that, recall, we used in the introductory section to explain the importance of model uncertainty methods. In the Supporting information (Web Appendix C), we provide the details for estimating this curve, which differs substantially from Nikooienejad et al. (2020). In addition, caution must be placed in reporting model uncertainty estimates of regression parameters. Unavoidable, the posterior distribution of β_j is a mixture of continuous densities and a point mass at zero (with probability $1 - p_j$, i.e., one minus its posterior inclusion probability). In these circumstances, conventional summaries (like the posterior mean) are inappropriate, and one has to resort to alternative reports that acknowledge this singularity of the posterior distribution, as we detail in the context of an actual study in Section 5.

3 | CONVENTIONAL PRIORS

The posterior distribution depends on the prior over the model space, $p(M_\gamma)$, and the priors for the regression parameters under each model $\pi_\gamma(\boldsymbol{\beta}_\gamma)$. In this section, we study the assignment of these from an objective perspective.

3.1 | Prior distribution over the model space

The prior over the model space is an important issue. The default objective choices for this distribution are the uniform, $p(M_\gamma) = 1/2^p$, or the hierarchical uniform prior discussed by Scott and Berger (2010):

$$p(M_\gamma) = \frac{1}{p+1} \binom{p}{k_\gamma}^{-1}. \tag{8}$$

This prior assigns the same probability to models of the same size k_γ. Our recommendation is the latter, as it considers the multiplicity of comparisons, as has been well argued by Scott and Berger (2010). Nevertheless, for specific scenarios, other choices for $p(M_\gamma)$ may be more appealing by trying, for example, to force sparsity (Castillo et al., 2015).

3.2 | Prior for regression parameters

In this section, we develop and rationalize the prior distributions $\pi_\gamma(\boldsymbol{\beta}_\gamma)$ that we propose. As the priors on the model's parameters are assigned conditionally to model M_γ, we drop the sub-index γ in all our expressions in this section. Hence, $\boldsymbol{\beta}_\gamma$ is simply notated as $\boldsymbol{\beta}$; $\boldsymbol{X}_\gamma$ as $\mathbf{X}$; k_γ as k and so on.

In model selection, objective approaches to the specification of $\pi(\boldsymbol{\beta})$ cannot rely on improper or vague proper priors (i.e., those with an arbitrarily large variance/scale). It is well-known that either of these would lead to indeterminate Bayes factors (Berger & Pericchi (2001); Cabras et al. (2014, 2015)).

Our proposal follows the tradition of conventional priors where $\pi(\boldsymbol{\beta}) = N_k(\mathbf{0}, g\boldsymbol{\Sigma})$, that is, a zero-mean normal distribution with variance $g\boldsymbol{\Sigma}$. The hyper-parameter g can be assigned a density, as done before by several authors (see, e.g., Bayarri et al., 2012; Li & Clyde, 2018). Nevertheless, in the simulated problems and real applications that we have analyzed, we find that results are, to a great extent, quite robust to this choice. In the context of large model spaces, the need for a computationally feasible procedure is necessary; thus, we recommend considering $g = 1$ fixed. Nevertheless, results are highly dependent on the prior variance, so a sensible objective assignment of the prior should be based on a conscientious specification of $\boldsymbol{\Sigma}$.

The route we follow to obtain $\boldsymbol{\Sigma}$ mimics the main ideas in the literature of conventional priors and is strongly guided by the expected Fisher information matrix evaluated in the null model. Details of the derivation are provided in the Supporting information (Web Appendix A). Notably, and despite the particular nature of the Cox model, we obtain a weighted version of the precision matrix of the covariates, a distinctive ingredient in the conventional priors in the normal linear model. In particular, we propose

$$\boldsymbol{\Sigma} = \left(\sum_{i=1}^{n} w_i^\star (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_w)^\top \right)^{-1}, \tag{9}$$

where $\bar{\boldsymbol{x}}_w = \sum_{i=1}^n w_i^\star \boldsymbol{x}_i$, $w_i^\star = w_i / \sum_{i=1}^n w_i$ and

$$\begin{aligned} w_i &= p(\delta_i = 1 \mid M_0) + p(\delta_i = 0 \mid M_0)\, H_0(c_i) \\ &= 1 - \mathrm{e}^{-H_0(c_i)} + \mathrm{e}^{-H_0(c_i)} H_0(c_i). \end{aligned} \tag{10}$$

The cumulative risk, $H_0(y)$, is unknown, and under the null model, where $\boldsymbol{\beta} = 0$, it can be estimated using the Breslow estimator or other parametric estimators based on the Weibull model. This strategy is also used and mentioned in Section 9 in Castellanos et al. (2021), where common parameters are estimated for all models, under the null model, to make calculations faster. Both strategies for estimating the baseline hazard function under the null model in a non-parametric model or under the Weibull model have been analyzed in the examples and the simulation studies, providing similar results.

Remarkable, a particular case of $\boldsymbol{\Sigma}$ is the precision matrix:

$$\left(\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})^{\top}/n\right)^{-1}, \tag{11}$$

which is the prior covariance matrix used unanimously in the definition of g-priors in the normal linear model (see, for instance, Zellner and Siow 1980; Liang et al. 2008 or Bayarri et al. 2012). This particular case arises when all censoring times are equal, and hence there is no possibility of distinguishing (a priori) which observational units are more or less informative.

When the c_is vary, $\boldsymbol{\Sigma}$ may substantially differ from Equation (11). Matrix, $\boldsymbol{\Sigma}$, underweights the covariate values of an individual with a short censoring time (thus with a small probability of being uncensored) and favors the covariate values of individuals with large censoring times. The trade-off between the two will affect how the values of the covariates participate in the prior covariance matrix. This convenient adaptive feature can be seen in the extreme situation formalized in the following result, whose proof is contained in Web Appendix B.

Lemma 1. ***Polarized censoring times***

Consider the situation where the sample units are highly polarized, with some having minimal censoring times compared to the rest. A canonical case of this is when $\boldsymbol{c} = (c_u, \overset{n_u}{\ldots}, c_u, c_c, \overset{n_c}{\ldots}, c_c)$. *In this case, as* c_c *decreases,*

$$\boldsymbol{\Sigma} \underset{c_c\to 0}{\to} \left(\sum_{i=1}^{n_u}(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_u)(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_u)^{\top}/n_u\right)^{-1}, \tag{12}$$

it converges to the (inverse of) the covariance matrix considering only the uncensored observations; here, $\bar{\boldsymbol{x}}_u$ *denotes the mean of the variables over the uncensored observations.*

Castellanos et al. (2021) have already argued in favor of a prior covariance matrix with an "adaptive" behavior to the amount of (a priori expected) information provided for each subject. Their main argument was that sampling units with short censoring times (hence with a negligible impact on the likelihood function) have the potential to modify prior covariance matrices in such a way that the results are highly distorted. In summary, this is an undesirable consequence of a conflict between the likelihood of individuals contributing different amounts of information (depending on whether or not they are censored) and a prior to which all individuals would equally contribute through the design matrix. In the context of the lognormal model for Y, Castellanos et al. (2021) illustrated this effect through a hypothetical situation using a simulated dataset, where purposely censored and uncensored subjects differ substantially in the values of their corresponding covariates.

TABLE 1 Inclusion probabilities of covariates in the primary biliary cholangitis (PBC) dataset with non local-priors (BVSNLP) and conventional priors (BVSCP)

	BVSCP	BVSNLP	BIC
Age	0.948	0.915	1.00
Albumin	0.944	0.919	0.99
Alk.phos	0.072	0.000	0.04
Ascites	0.090	0.002	0.07
Ast	0.170	0.038	0.19
Bili	0.999	1.000	1.00
Chol	0.054	0.000	0.02
Copper	0.296	0.066	0.40
Edema	0.613	0.360	0.78
Hepato	0.099	0.005	0.06
Platelet	0.115	0.007	0.07
Protime	0.686	0.574	0.77
Spiders	0.062	0.000	0.02
Stage	0.470	0.222	0.49
Trig	0.133	0.004	0.15

Surprisingly, this situation is not unusual in real datasets. For instance, in the PBC dataset used in the introduction, some of the explanatory variables exhibit such differential variability, as Volinsky and Raftery (2000) pointed out. Specifically, in covariate *ascites* the variance of all subjects is twice that of uncensored observations and is 1.6 times larger in *edema*. There is a difference in the prior variance if it results from underweights censored observations (as the likelihood function does) or if it results from all subjects equally participating in constructing the prior variance: Bayes factors and posterior probabilities will be sensitive to this choice. In Table 1, we provide the posterior inclusion probabilities of the main effects of the explanatory variables with our approach, that is, $\pi(\boldsymbol{\beta}) = N_k(\mathbf{0}, g\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ given in Equation (9) (labeled BVSCP) and $g = 1$. For comparative purposes, we also include results with the method by Nikooienejad et al. (2020) as implemented by the authors in the R package BVSNLP with the function bvs. The latter methodology, which we label BVSNLP throughout the paper, uses all the values of covariates to determine the prior variance without distinction between censored and uncensored observations. In both cases, BVSCP and BVSNLP, the prior over the model space adopted is Equation (8).

From Table 1, we can see that underweighting censored observations, as is done with BVSCP, promotes a more complex model than BVSNLP. The inclusion probabilities for most covariates are higher in BVSCP than in BVSNLP. At the same time, the relative evidence among the covariates is kept constant across the methods BVSCP versus BVSNLP (see, for instance, the case of *bili* covariate).

An intriguing case is that of *edema*, with a posterior inclusion probability of 0.613 with BVSCP and barely 0.360 with BVSNLP. Recall, *edema* is one of the variables for which its variability differs significantly between the uncensored subjects and the whole sample. This result is in line with the example highlighted in the motivating dataset in Castellanos et al. (2021). Of course, it is unknown if *edema* belongs to the "true" model, although previous studies suggest it. Particularly, the methodology in Volinsky and Raftery (2000) implemented in the R package BMA with the function bic.surv and that we label BIC gives *edema* an inclusion probability of 0.78 (see Table 1). We will return to this dataset in Section 5.

4 | NUMERICAL METHODS

In this section, we provide practical recommendations to implement the proposed approach.

4.1 | Approximating the marginal

To compute the integral that defines $m(\mathbf{y})$ in Equation (5) and to speed up the calculations at the cost of using approximations, we massively employ the Laplace approximation, as also done in Bové and Held (2011) and Nikooienejad et al. (2020). It results in

$$m(\mathbf{y}) \approx L_p(M, \widehat{\boldsymbol{\beta}})\pi(\widehat{\boldsymbol{\beta}})(2\pi)^{k/2}|H_{\widehat{\boldsymbol{\beta}}}|^{-1/2}, \tag{13}$$

where $\widehat{\boldsymbol{\beta}}$ is the maximum a posteriori estimation of $\boldsymbol{\beta}$ and $H_{\widehat{\boldsymbol{\beta}}}$ is the Hessian of the negative of the log posterior, that is, the Hessian of $-\log L_p(M, \widehat{\boldsymbol{\beta}}) - \log \pi(\boldsymbol{\beta})$.

To compute Equation (13), we consider the minimization algorithm implemented in the optim function in R.

4.2 | Model search

When considering large values of p, in particular for $p > 30$, exhaustive enumeration is unfeasible. In this section, we demonstrate the ability of a Gibbs sampling algorithm (George & McCulloch, 1997) to perform the model search adequately.

The Gibbs sampling algorithm is initialized in a model $\boldsymbol{\gamma}_{(0)} = (\gamma_{1(0)}, \gamma_{2(0)}, \dots, \gamma_{p(0)})$ with associated (Laplace) approximated Bayes factor, $B_{\gamma_{(0)}}$, then repeating, for $i = 1, \dots, N + B$, where $B > 1$ is the burn-in period:

- ◇ *Step* j : $1 \le j \le p$. Consider the candidate model $\boldsymbol{\gamma}_* = (\gamma_{1(i-1)}, \dots, 1 - \gamma_{j(i-1)}, \dots, \gamma_{p(i-1)})$. If this model has been sampled before, use its already computed B_{γ_*}; otherwise, compute it and save it. Compute the selection probability:

$$r_{i,j} = \frac{B_{\gamma_*} p(M_{\gamma_*})}{B_{\gamma_*} p(M_{\gamma_*}) + B_{\gamma_{(i-1)}} p(M_{\gamma_{(i-1)}})}, \tag{14}$$

  and with probability $\min\{r_{i,j}, 1\}$ re-define $\boldsymbol{\gamma}_{(i-1)} = \boldsymbol{\gamma}_*$.
- ◇ *Final step*. Define and save $\boldsymbol{\gamma}_{(i)} = \boldsymbol{\gamma}_{(i-1)}$ and $B_{\gamma_{(i)}} p(M_{\gamma_{(i)}})$, a quantity that is proportional to the posterior probability of this model.

Discarding the first B iterations, we end up with N samples (models) simulated (possibly with repetitions) from the posterior distribution. To estimate the inclusion probability of variable x_j, we use the Rao-Blackwellized estimator proposed in Ghosh and Clyde (2011), and Gelfand and Smith (1990). These estimates are constructed by replacing the quantity of interest in the Monte Carlo estimator with its expectation, given other components in the sampler. In our case:

$$\hat{\pi}^{\mathrm{RB}}_{N,j} = \frac{1}{N} \sum_{i=1}^{N} (1 - \gamma_{j(i-1)}) r_{i,j} + \gamma_{j(i-1)}(1 - r_{i,j}). \tag{15}$$

As we can see in Figure 2, Rao-Blackwellized estimates of inclusion probabilities for the PBC data stabilize in a small number of Gibbs iterations ($N \approx$ 200–300).

5 | CASE STUDY: PRIMARY BILIARY CHOLANGITIS DATA

The Mayo Clinic conducted a trial on the liver's PBC between 1974 and 1984. The dataset is distributed in the survival package, and we consider the $p = 15$ covariates described in the Web Appendix F. The Gibbs sampling algorithm has approximated posterior model probabilities

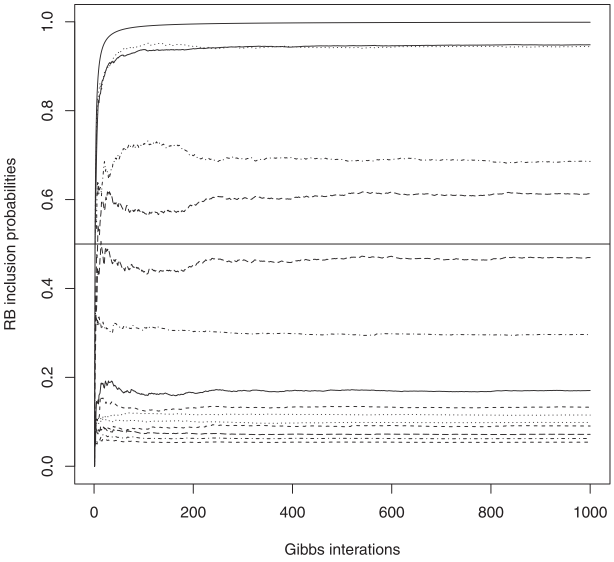

FIGURE 2 Rao-Blackwellized estimates of inclusion probabilities along Gibbs iterations for the primary biliary cholangitis (PBC) data

in $N = 20,000$ iterations. A small number of steps allows the inclusion probabilities to be estimated adequately, as shown in Figure 2.

Table 2 shows the five most probable models, together with their posterior probabilities, which only add up to 26%. The ten most probable models add up to 37% of the posterior model probability—73% the top 100. This result reflects the above-mentioned fact; that there is a lot of model uncertainty in this dataset.

Figure 3 contains the approximated model-averaged posterior distribution for the most important covariates with higher inclusion probabilities. The dark gray bar represents the probability of no effect for each covariate in each histogram. Given an effect, the rest of the histogram area represents its posterior distribution. For example, *bili* has a positive risk effect, increasing the risk of death, followed by *age* and *albumin*, the former increasing the risk and the latter decreasing it. The following variable with some possible effect is *protime* with an approximated probability of being an explanatory factor of 0.68. *edema* also has a higher probability of being a risk factor for death, around 0.61 (0.39 probability of no effect). Finally, *stage* has a higher probability of no effect, around 0.53.

To evaluate the underestimation of model uncertainty made when using only the HPM, Figure 4 presents the histograms for `bili` considering BMA versus the HPM. The posterior distribution based on only one model is more concentrated, as expected, and hides the uncertainty about the models, which is reflected in the effect of the `bili` covariate.

We finish this section by completing the comparison among the competing procedures in the context of the PBC dataset. Regarding model uncertainty estimations, a leading aspect that governs the summaries is the inclusion probability (the dark bar in the histograms in Figure 3). We already showed, in Table 1 that these probabilities differed, particularly between BVSCP and BVSNLP, and the reason is how the sampling information from the covariates is incorporated into the priors. Nevertheless, apart from the priors, BVSCP and BVSNLP disagree in other fundamental aspects of how model uncertainty is handled, and the consequences are visible in the analysis of this dataset. In Table 2, we have collected comparing information about the most probable models in BVSCP. Remarkably, the five best models accumulate a probability (as measured by BVSCP) of barely 0.26 while, as obtained with BVSNLP, the posterior probability of the best two already exceeds 0.50. This result manifests the overestimation of probabilities caused by summaries obtained from a heuristic searching method (aimed at discovering good models that need re-normalization). BIC prefers more complex models, a tendency we already observed in the simulation study that leaves an undesirable behavior with a more considerable risk of promoting false positives.

6 | DISCUSSION

We have addressed the problem of model uncertainty in the context of Cox regression, perhaps the most popular statistical procedure in survival analysis. Our proposal builds on intertwined theoretical and numerical

TABLE 2 The table reports the order of the five most probable models with each method: BVSNLP, BVSCP, and BIC, jointly with the posterior probability in each case

model:	**BVSCP**		**BVSNLP**		**BIC**	
{age, bili, albumin} $\cup$	**Order**	$p(M_\gamma \mid y)$	**Order**	$p(M_\gamma \mid y)$	**Order**	$p(M_\gamma \mid y)$
{protime}	1	0.082	1	0.411	> 5	< 0.04
{edema, protime}	2	0.062	> 5	< 0.05	1	0.071
{edema, stage}	3	0.048	4	0.053	4	0.042
{edema}	4	0.035	2	0.144	> 5	< 0.04
{edema, protime, stage}	5	0.032	> 5	< 0.05	3	0.049

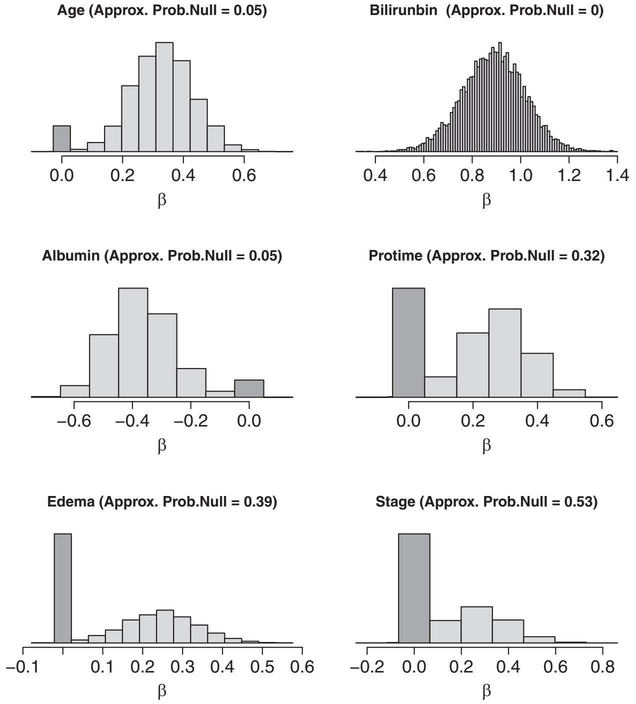

FIGURE 3 Primary biliary cholangitis (PBC) dataset. Model-averaged posterior distributions of the regression coefficients for some potential covariates. The dark gray area represents the probability of no effect, while the light gray area represents the approximated probability distribution given an effect. The title appears as the approximated probability of no effect (approximately the complementary of the inclusion probability for each covariate).

procedures, which are novel to the problem under study and which, when put together, result in a solid and practical methodological tool for practitioners. In particular, we have derived a new prior distribution for the specific model parameters (based on the expected information matrix, called conventional). Furthermore, we have argued decisively in favor of an assignment that controls for multiplicity in terms of what the prior to each subset of explanatory variables refers to. The resulting prior scheme is objective and fully automatic, so there is no need to tune any additional parameters.

The model space is vast, and even if p, the number of regressors, were modest, exhaustive enumeration of all models is unfeasible. This difficulty has previously been addressed either with a severe pruning of the space (the case of Hoeting et al., 1999a) or with optimization-based procedures aimed at finding the best models (as proposed by Nikooienejad et al., 2020). The first solution only accounts for the uncertainty provided by a tiny portion of the model space, compromising the essential pillar of model uncertainty. The second possibility leads to biased results, as documented by Garcia-Donato and Martinez-Beneito (2013), simply because more probable models are over-represented (and the contrary with less likely ones). We propose implementing a Gibbs sampling algorithm in combination with a Rao-Blackwellization estimation of probabilities. A probabilistic sampling of the model space ensures that the obtained estimations reflect the variability present in the problem. In the simulations described in the Supporting information, we have shown the ability of the numerical method to obtain pretty accurate summaries of the posterior probability, even using very few iterations.

We finally discuss three interesting complementary questions raised in the reviewing process of this work and for which we thank the referees. A first question is whether the order used to sample the components in $\boldsymbol{\gamma}$ matters. Sampling theory states that, after the burning period, we will be simulating from the posterior distribution no matter the order used in the Gibbs sampling. We have checked this property numerically, running simulations of our numerical study with random permutations of the $p = 1,000$ components of $\boldsymbol{\gamma}$, obtaining identical results. A second observation is about the feasibility and goodness of the method when handling problems with higher p. In Web Appendix G of the Supporting information, we have repeated the simulation study but with $p = 10,000$. We observe similar patterns as those highlighted with $p = 1,000$, concluding that our method behaves satisfactorily in high-dimensional settings. Of course, the procedure's feasibility in ultrahigh problems is still an open problem. We are currently working on scalable implementations of the Gibbs algorithm that could, in principle, handle these challenging situations. Another interesting question was what would happen if the actual data generative model is not in the list of candidate models. For instance, the hypothesis of proportional risks would not hold. This context is called the $\mathcal{M}$-open perspective in the literature. The theory states that, asymptotically, the model which is closest to the true one (in terms of Kullback–Liebler discrepancy) will be given maximum posterior probability (Berk, 1966; Dmochowski, 1996). Nevertheless, although these results condition what we expect in a finite sample size, their practical implications are difficult to envisage (and are rarely examined). To contribute to this discussion, in Web Appendix H, we have launched the counterpart of Scenario 2 contained in the simulation study in Web Appendix D, but where the actual model does not satisfy the proportionality assumption. In particular, we have simulated accelerated failure times with normal (Scenario 5), and Cauchy (Scenario 6) distributed errors. The interpretation of results is far from being straightforward, and the question of which Cox regression model is "closest" to the true data-generative model emerges. More specific research is needed in this direction, but the preliminary conclusions are undoubtedly positive. Cox models endorsed by larger probabilities seem to

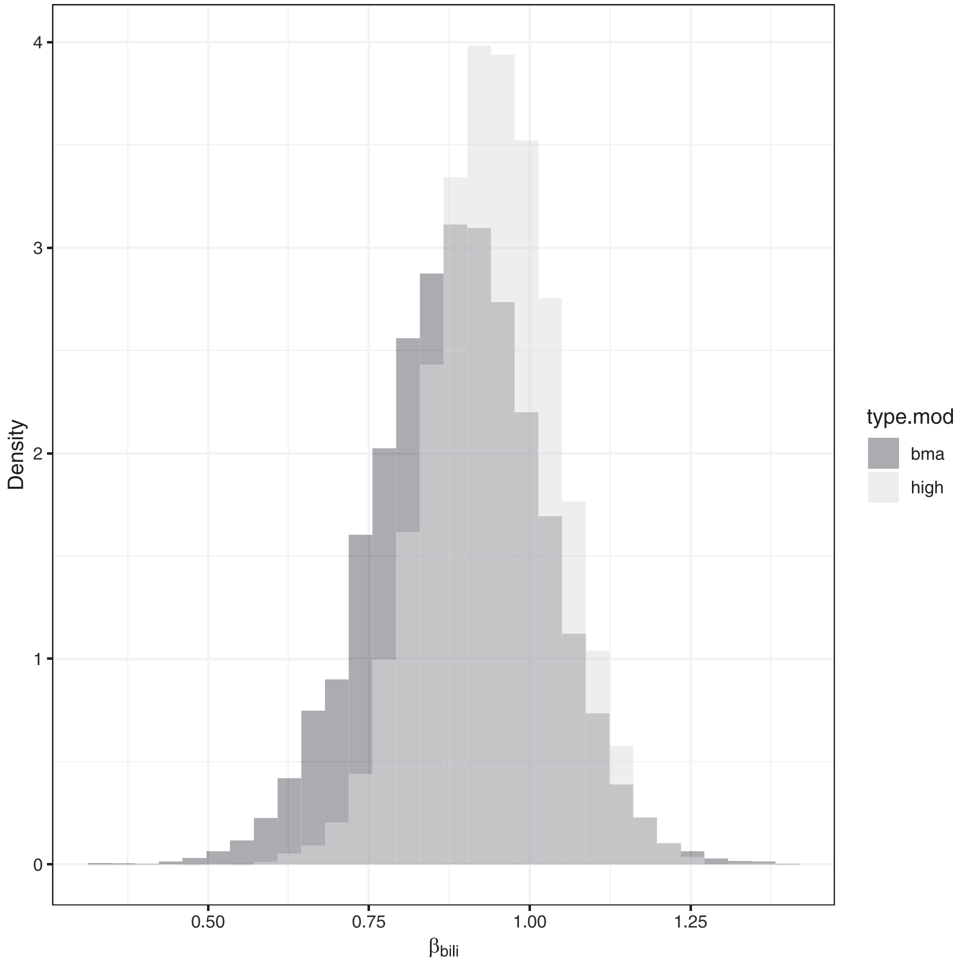

FIGURE 4 Posterior probability distribution for *Bilirunbin* (`bili`), considering model averaged (dark gray) or the highest probability model (light gray)

reasonably mimic the true covariate structure (which are the true explanatory and which are not). The third point relates to the advantages of model uncertainty methods, like the ones here developed, in relation to produce results with good frequentist coverages. In the simulation studies, we show that, when compared with credible intervals using a single model (the HPM), methods that account for model uncertainty have superior performance in terms of coverage.

ACKNOWLEDGMENTS

The work of all authors has been partially supported by the MICINN project PID2019-104790GB-I00. The authors thank two referees and an Associate Editor for a careful reading of our manuscript, and the many useful comments received that have largely enriched this research.

DATA AVAILABILITY STATEMENT

Web Appendix F contains a description of the data and access to R code implementing the methodology for the case study in Section 5.

ORCID

Gonzalo García-Donato https://orcid.org/0000-0002-5642-0042
Stefano Cabras https://orcid.org/0000-0001-6690-8378
María Eugenia Castellanos https://orcid.org/0000-0001-7920-2307

REFERENCES

Bailey, K.R. (1983) The asymptotic joint distribution of regression and survival parameter estimates in the Cox regression model. *The Annals of Statistics*, 11(1), 39–48.

Barbieri, M.M. & Berger, J.O. (2004) Optimal predictive model selection. *The Annals of Statistics*, 32(3), 870–897.

Barbieri, M.M., Berger, J.O., George, E.I. & Ročková, V. (2021) The median probability model and correlated variables. *Bayesian Analysis*, 16(4): 1085–1112. DOI: 10.1214/20-BA1249

Bayarri, M., Berger, J., Forte, A. & García-Donato, G. (2012) Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3), 1550–1577.

Berger, J.O. & Pericchi, L.R. (2001) Objective Bayesian methods for model selection: Introduction and comparison. In: Lahiri, P. (Ed.) *Model selection*, vol. 38, Institute of Mathematical Statistics, pp. 135–207.

Berk, R. (1966) Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics*, 37, 51–58.

Bové, D.S. & Held, L. (2011) Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3), 387–410.

Cabras, S., Castellanos, M.E. & 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.

Cabras, S., Castellanos, M.E. & Perra, S. (2015) A new minimal training sample scheme for intrinsic Bayes factors in censored data. *Computational Statistics & Data Analysis*, 81, 52–63.

Castellanos, M.E., García-Donato, G. & Cabras, S. (2021) A model selection approach for variable selection with censored data. *Bayesian Analysis*, 16(1), 271–300.

Castillo, I., Schmidt-Hieber, J. & van der Vaart, A. (2015) Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5), 1986–2018.

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

Dmochowski, J. (1996) Intrinsic priors via Kullback–Liebler geometry. In: Bernardo, J., Berger, J., Dawid, A. & Smith, A. (Eds.) *Bayesian statistics*, vol. 5. London: Oxford University Press, pp. 543–550.

Etz, A. & Wagenmakers, E.-J. (2017) J.B.S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, 32(2), 313–329.

Fleming, T.R. & Harrington, D.P. (2011) *Counting processes and survival analysis*, vol. 169. John Wiley & Sons.

Garcia-Donato, G. & 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.

Gelfand, A.E. & Smith, A. F.M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410), 398–409.

George, E.I. & McCulloch, R.E. (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, 7, 339–373.

Ghosh, J. & Clyde, M.A. (2011) Rao-Blackwellization for Bayesian variable selection and model averaging in linear and binary regression: a novel data augmentation approach. *Journal of the American Statistical Association*, 106(495), 1041–1052.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999a) Bayesian model averaging: a tutorial. *Statistical Science*, 14(4), pp. 382–401.

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. (1999b) Bayesian model averaging: a tutorial. *Statistical Science*, 14(4), 382–401.

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

Johansen, S. (1983) An extension of Cox's regression model. *International Statistical Review/Revue Internationale de Statistique*, 51(2), 165–174.

Johnson, V.E. & 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.

Johnson, V.E. & Rossell, D. (2012) Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498), 649–660.

Kalbfleisch, J.D. (1978) Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2), 214–221.

Li, Y. & Clyde, M.A. (2018) Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524), 1828–1845.

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

Murphy, S.A. & Vaart, A.W.V.D. (2000) On profile likelihood. *Journal of the American Statistical Association*, 95(450), 449–465.

Nikooienejad, A., Wang, W. & Johnson, V.E. (2020) Bayesian variable selection for survival data using inverse moment priors. *The Annals of Applied Statistics*, 14(2), 809–828.

Raftery, A.E., Madigan, D. & Hoeting, J. (1997) Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92, 179–191.

Robert, C.P., Chopin, N. & Rousseau, J. (2009) Harold Jeffreys' theory of probability revisited. *Statistical Science*, 24(2), 141–172.

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

Sinha, D., Ibrahim, J.G. & Chen, M.H. (2003) A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90(3), 629–641.

Steel, M.F.J. (2020) Model averaging and its use in economics. *Journal of Economic Literature*, 58(3), 644–719.

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

Zellner, A. & Siow, A. (1980) Posterior odds ratios for selected regression hypotheses. In: Bernardo, J.M., Berger, J.O., Dawid, A.P. & Smith, A.F.M. (Eds.) *Bayesian statistics 1*, vol. 31. Berlin: Springer, pp. 585–603. doi:10.1007/BF02888369.

SUPPORTING INFORMATION

Web Appendices referenced in Sections 1–3, 5 and 6 are available with this paper at the Biometrics website on Wiley Online Library. These include all technical details, additional simulation results, data description and access to R code implementing the methodology for the case study in Section 5.

How to cite this article: García-Donato, G., Cabras, S. & Castellanos, M.E. (2023) Model uncertainty quantification in Cox regression. *Biometrics*, 79, 1726–1736. https://doi.org/10.1111/biom.13823