Statistical Science
2023, Vol. 38, No. 1, 13–29
https://doi.org/10.1214/21-STS846

Additive Bayesian Variable Selection under Censoring and Misspecification

David Rossell and Francisco Javier Rubio

Abstract. We discuss the role of misspecification and censoring on Bayesian model selection in the contexts of right-censored survival and concave log-likelihood regression. Misspecification includes wrongly assuming the censoring mechanism to be noninformative. Emphasis is placed on additive accelerated failure time, Cox proportional hazards and probit models. We offer a theoretical treatment that includes local and nonlocal priors, and a general nonlinear effect decomposition to improve power-sparsity trade-offs. We discuss a fundamental question: what solution can one hope to obtain when (inevitably) models are misspecified, and how to interpret it? Asymptotically, covariates that do not have predictive power for neither the outcome nor (for survival data) censoring times, in the sense of reducing a likelihood-associated loss, are discarded. Misspecification and censoring have an asymptotically negligible effect on false positives, but their impact on power is exponential. We show that it can be advantageous to consider simple models that are computationally practical yet attain good power to detect potentially complex effects, including the use of finite-dimensional basis to detect truly nonparametric effects. We also discuss algorithms to capitalize on sufficient statistics and fast likelihood approximations for Gaussian-based survival and binary models.

Key words and phrases: Additive regression, generalized additive model, misspecification, model selection, survival.

1. INTRODUCTION

Determining what covariates have an effect on a survival (time-to-event) outcome is an important task in many fields, including Biomedicine, Economics and Engineering. For interpretability and computational convenience it is common to use parametric and semiparametric models such as Cox proportional hazards [8] or accelerated failure time (AFT) regression for survival outcomes, possibly with nonlinear additive effects. The proportional hazards model assumes that covariates have a multiplicative effect on the baseline hazard, whereas in AFT models covariates drive the mean of the logarithmic (or other monotonic transform) times-to-event. These models can be combined with Bayesian model selection to provide a powerful mechanism to select variables, enforce sparsity and quantify uncertainty. However, the precise consequences of model misspecification and censoring are not sufficiently understood. By misspecification we mean that the data are truly generated by a distribution outside the considered class. For instance, one may fail to record truly relevant covariates or represent their effects imperfectly, for example, when using a Cox model but the true covariate effects on the hazard are nonproportional. This issue can be addressed by enriching the model, for example, via nonparametric or time-dependent effects. Then, the potential concerns are that the larger number of parameters can adversely affect inference, unless the sample size is large enough, and that computation can be costlier. Censoring is also important. First, it reduces the effective sample size. Second, wrongly assuming the censoring mechanism to be noninformative, that is, independent of the outcome conditionally on covariates, may affect the selected model, even asymptotically.

Our goal is to help understand the consequences of three important issues on model selection: misspecification, censoring, and trade-offs when including nonlinear effects. We first consider that the data analyst assumed a nonlinear additive AFT model, or an additive Cox model, but data are truly generated by a different probability dis-

David Rossell is Associate Professor, Department of Economics and Business, Pompeu Fabra University, Barcelona, Spain (e-mail: david.rossell@upf.edu). Francisco Javier Rubio is Assistant Professor, Department of Statistical Science, University College London, London, United Kingdom (e-mail: f.j.rubio@ucl.ac.uk).

tribution F_0. We also consider probit regression, which can be formulated as a particular case of the Normal AFT model, and more general concave log-likelihood regression, which provides a unifying framework for the models we consider here.

There are many data analysis methods for survival outcomes, along with theory for well-specified models and empirical results suggesting potential issues under misspecification, but their implications for model selection have not been described in sufficient detail. We first review results on the behavior of misspecified AFT and proportional hazard models, and subsequently discuss some model selection methods for survival data. Although both models have similar asymptotic properties, and which model is more appropriate depends on the data at hand, AFT inference has been argued to be more robust and to better preserve interpretability under misspecification. More precisely, the maximum likelihood estimators under misspecified Cox and AFT models have comparable limiting distributions if censoring is absent or independent of covariates [51, 58], but not so under covariate-dependent censoring [49]. Covariate-dependent censoring also affects frequentist hypothesis tests. In misspecified Cox models, it can lead to a substantial type I error inflation [9]. In misspecified AFT models, the power of the tests may be affected, but simple strategies to control the type I error are available [16, 21, 49].

Another situation where both models behave differently is when omitting truly active covariates, for example, because these were not recorded. A proportional hazards model with omitted variables tends to underestimate covariate effects, even for a treatment of interest that is uncorrelated with other covariates [25, 51]. Further, even if the data-generating truth has proportional hazards, the marginal model conditioning only on the observed covariates does not (except in positive stable distributions, [19]). In contrast, if the data-generating truth is an AFT model and one omits relevant covariates, the unaccounted variability is subsumed into the error term, and regression parameters remain interpretable as averaged effects across the population [21]. Note that omitting covariates is intimately connected to incorporating a covariate but misspecifying its effect: using a linear or finite-dimensional effect can be seen as omitting a subset of the columns of the basis defining a truly nonparametric effect. Thus our discussion on omitted variables applies directly to misspecifying covariate effects. To summarize, censoring and misspecification have nontrivial effects on estimation and hypothesis testing.

We now review some model selection methods for survival data, discussing the extent to which they considered misspecification. [20, 26, 48] proposed likelihood penalties for Cox and semiparametric AFT models, and [53] for broader generalized hazards models. Most of this work focused on linear covariate effects, computation and proving consistency under covariate-independent censoring. There are however empirical results on the effect of misspecification, for example, [60] showed in simulations an increase in false positives of the Cox–LASSO method [52] when data truly arise from an AFT model. There are also many Bayesian variable selection methods for survival data. [11] and [46] proposed shrinkage priors for the Cox and AFT models and assessed performance via simulations where the model was well specified. [22] studied Bayesian model selection for the Cox model, [10] for the so-called additive hazards model, and [33] for the Cox model under nonlocal priors [23, 24]. See [28] for a review, with a focus on the Cox model. While interesting, these Bayesian proposals do not consider misspecification. [39] did study misspecified Bayesian linear regression, showing that misspecification often reduces the power to detect active variables, but did not consider censoring.

We summarize our main messages. We show that, under mild assumptions, Bayesian model selection asymptotically discards covariates that do not predict neither the outcome nor the censoring times. By *predict*, we refer to increasing the expectation of the log-likelihood function. For any fully specified model, said expectation is a weighted average of a reward for assigning a high probability to the observed censoring time in individuals that are censored, and a reward for predicting survival times accurately in uncensored individuals (e.g., mean squared error, for Normal AFT models). For the partially specified Cox model, the reward is for assigning a high hazard to individuals who experienced the event, relative to other individuals at risk. We discuss that both censoring and wrongly specifying covariate effects have an exponential effect in power, but that asymptotically neither leads to false positive inflation. We also develop a novel nonlinear effect decomposition to ameliorate the power drop, and study the consequences of using finite basis to describe covariate effects, a practical strategy to speed up computations when one considers many models. For concreteness, we outline a formulation based on a novel combination of nonlocal priors [23] and group-Zellner priors that induce group-level sparsity for nonlinear effects. As a technical contribution, we prove the asymptotic validity of Laplace approximations to Bayes factors for concave log-likelihoods under minimal conditions, allowing for misspecification, which provides a simple basis to study Bayes factors that covers all models considered in this paper. We also provide software (R package `mombf`).

The outline is as follows. Section 2 discusses the likelihood for AFT and Cox models, priors and a nonlinear effect decomposition aimed at improving power. Section 3 discusses known and novel results on asymptotic normality and Bayes factor rates, and how to interpret

the Bayesian model selection solution under misspecification. Similar results are obtained for general concave log-likelihoods; see the Supplementary Material [40]. See also there for a known but seemingly unexploited result in the literature, that probit models are a particular case of the Normal AFT model. Section 4 discusses the relative computational convenience of AFT vs. Cox models related to the use of sufficient statistics. It also discusses an approximation to the Normal log-distribution function and derivatives that significantly increases speed and accuracy, and may have some independent interest, and simple model exploration strategies. Section 5 illustrates the effect of misspecification and censoring in simulations and cancer datasets, practical power-sparsity tradeoffs, and the use of finite-dimensional nonlinear basis. Section 6 concludes. The Supplementary Material contains derivations related to the likelihood, priors and their derivatives, and prior elicitation. It also offers detailed discussions on computational algorithms, including a novel approximation to Normal log-distribution functions that may have some independent interest. Finally, it contains all proofs for our main results, additional propositions for the AFT model with Laplace errors and probit models, the asymptotic validity of Laplace approximations to integrated likelihoods, and empirical results that supplement those in the main paper.

2. FORMULATION

Our discussion focuses on survival data, but see the Supplementary Material for binary regression and more general concave log-likelihoods. Section 2.1 sets notation, reviews the AFT and proportional hazards models, their being a particular cases of the generalized hazards structure, and a nonlinear effects decomposition to improve interpretability and power. Section 2.2 embeds the problem within a Bayesian model selection framework. Section 2.3 introduces prior distributions that can accommodate group and hierarchical constraints, and Section 2.4 suggests default prior parameter values.

2.1 Likelihood

Let us introduce the notation. Suppose that one is interested in studying the dependence of a survival (or time-to-event) outcome $o_i \in \mathbb{R}_+$ on a covariate vector $x_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, for individuals $i = 1, \ldots, n$. Suppose that there are right-censoring times $c_i \in \mathbb{R}_+$, such that one only observes the outcome for uncensored individuals, that is, those for which $o_i \leq c_i$. Denote by $u_i = \mathrm{I}(o_i < c_i)$ the indicator that observation i is uncensored, $y_i = \min\{\log(o_i), \log(c_i)\}$ the observed log-times, $y = (y_1, \ldots, y_n)$, $u = (u_1, \ldots, u_n)$, and the number of uncensored individuals $n_o = \sum_{i=1}^n u_i$.

We review two popular models for survival data, the AFT and Cox models, and discuss a strategy to decompose nonlinear effects. The AFT model postulates

$$\log(o_i) = \sum_{j=1}^{p} g_j(x_{ij}) + \epsilon_i,$$

where $g_j : \mathbb{R} \to \mathbb{R}$ and ϵ_i are independent across $i = 1, \ldots, n$ with mean $E(\epsilon_i) = 0$ and variance $V(\epsilon_i) = \sigma^2$ (assumed finite). Typically, g_j is expressed in terms of an r-dimensional basis, for example, splines or wavelets [55]. For interpretability and to gain power (see Section 3.2) it is convenient to decompose g_j into a linear and a deviation-from-linearity components. To fix ideas, the cubic splines used in our examples consider

$$\log(o_i) = x_i^\top \beta + s_i^\top \delta + \epsilon_i, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$, $\delta^\top = (\delta_1^\top, \ldots, \delta_p^\top) \in \mathbb{R}^{rp}$ and $s_i^\top = (s_{i1}^\top, \ldots, s_{ip}^\top)$, where $s_{ij} \in \mathbb{R}^r$ is the projection of x_{ij} onto a cubic spline basis orthogonalized to x_{ij} (and the intercept). The idea is that $x_i^\top \beta$ captures linear effects, whereas $s_i^\top \delta$ captures deviations from linearity. Even if a covariate truly has a nonlinear effect, the linear term often captures a fraction of that effect using a single parameter, hence one can gain in power to detect its presence. Specifically we built s_{ij}, the ith row of the $n \times r$ matrix S_j, as follows. Let X_j and $\widetilde{S}_j$ have row i equal to $(1, x_{ij})$ and the cubic spline projection of x_{ij} (equi-distant knots), then $S_j = (I - X_j(X_j^\top X_j)^{-1} X_j^\top)\widetilde{S}_j$ is orthogonal to X_j. Denote by (X, S) the design matrix with $(x_i^\top, s_i^\top)$ in its ith row, and by (X_o, S_o) and (X_c, S_c) the submatrices with the rows for uncensored and censored individuals (respectively). This formulation contains partially linear models as particular cases, that is, when only some of the covariates are assumed to have a nonlinear effect). We denote the parameter space by $\Gamma \subset \mathbb{R}^{p(r+1)} \times \mathbb{R}^+$.

In survival analysis it is common to pose a model only for the survival times, such as (1). This is because the censoring is assumed to be noninformative given the covariates, and then the censoring distribution factors out of the likelihood function (Supplementary Material). The likelihood and partial likelihood used by the AFT and Cox models, which we review next, embed such a noninformativeness assumption. See Section 3 for a discussion on the consequences of this assumption not holding.

Regarding the likelihood associated to (1), consider the particular case where the errors are Gaussian. It is convenient to reparameterize $\alpha = \beta/\sigma$, $\kappa = \delta/\sigma$ and $\tau = 1/\sigma$, as then the log-likelihood is concave, provided the number of uncensored individuals is greater than the number of model parameters ($n_o \geq p + rp$) and that (X_o, S_o) has

full column rank [3, 47]. The log-likelihood is

$$\begin{aligned}(2)\qquad &\ell(\alpha,\kappa,\tau)\\ &= -\frac{n_o}{2}\log\Big(\frac{2\pi}{\tau^2}\Big) - \frac{1}{2}\sum_{u_i=1}(\tau y_i - x_i^\top\alpha - s_i^\top\kappa)^2\\ &\quad + \sum_{u_i=0}\log\{\Phi(x_i^\top\alpha + s_i^\top\kappa - \tau y_i)\};\end{aligned}$$

see the Supplementary Material for its gradient and hessian.

The Cox model instead assumes that the hazard function at time t takes the form

$$h_{PH}(t \mid x_i) = h_0(t)\exp\{x_i^\top\beta + s_i^\top\delta\},$$

where $h_0(\cdot)$ is a baseline hazard, typically estimated nonparametrically, and (β,δ) are estimated using the log partial likelihood [8]

$$(3)\quad \ell_p(\beta,\delta) = \sum_{u_i=1}\log\Big(\frac{\exp\{x_i^\top\beta + s_i^\top\delta\}}{\sum_{k\in\mathcal{R}(o_i)}\exp\{x_k^\top\beta + s_k^\top\delta\}}\Big),$$

where $\mathcal{R}(t) = \{i : o_i \geq t\}$ denotes the set of individuals at risk at time t.

To relate both models, (1) can be formulated in terms of the hazard function

$$h_{AFT}(t) = h_0\big(t\exp\{x_i^\top\beta + s_i^\top\delta\}\big)\exp\{x_i^\top\beta + s_i^\top\delta\}.$$

Both models are special cases of the generalized hazards structure [6]

$$(4)\quad h_{GH}(t) = h_0\big(t\exp\{x_i^\top\beta + s_i^\top\delta\}\big)\exp\{x_i^\top\theta + s_i^\top\xi\},$$

which we use in our examples to portray the behaviour of misspecified AFT and Cox models. Clearly, (4) contains the AFT model for $(\beta,\delta) = (\theta,\xi)$ and the proportional hazards model for $(\beta,\delta) = 0$.

2.2 Model Selection

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

$$\gamma_j = \begin{cases} 0, & \text{if } \beta_j = 0, \delta_j = 0,\\ 1, & \text{if } \beta_j \neq 0, \delta_j = 0,\\ 2, & \text{if } \beta_j \neq 0, \delta_j \neq 0,\end{cases}$$

corresponding to no effect, a linear and a nonlinear effect of each covariate $j = 1,\ldots,p$. That is, $\gamma = (\gamma_1,\ldots,\gamma_p)$ determines what covariates enter the model and their effect, and there are 3^p total models to consider. We remark that by nonlinear effect we refer to the specific effect coded by the chosen basis, for example, B-splines in our examples. One could extend the exercise by considering other types of nonlinear effects, for example, by adding a fourth possibility $\gamma_j = 3$ associated to a wavelet basis. Such basis would be orthogonalized to the linear term, as described after (1).

This formulation has two key ingredients. First, it decomposes effects into linear and deviation from linearity components, enforcing the hierarchical desiderata that the latter are only included if the linear terms are present. This decomposition is similar to the structured additive regression of [44], the main difference is that they do not test for exact $\beta_j = 0$, $\delta_j = 0$ and that they rely on a spectral decomposition that is less general than our simpler orthogonalization of (X_j, S_j). Our theory and results show that such decompositions improve the power to detect truly active effects. As discussed, this is because the option $\gamma_j = 1$ captures part of the effect of a variable with a single parameter. The second ingredient is considering the group inclusion of all nonlinear coefficients $\delta_j \in \mathbb{R}^r$. The motivation is that including individual entries in δ_j increases the probability of false positives, for example, if j truly had no effect there would be $2^r - 1$ subsets of S_j leading to including j.

Bayesian model selection proceeds as follows. Let $p_\gamma = \sum_{j=1}^p \mathrm{I}(\gamma_j \neq 0)$ be the number of active variables according to model γ, $s_\gamma = \sum_{j=1}^p \mathrm{I}(\gamma_j = 2)$ the number of nonlinear effects, and $d_\gamma = p_\gamma + rs_\gamma + 1$ the total number of parameters in γ for AFT models, and $d_\gamma = p_\gamma + rs_\gamma$ for Cox and probit models. (X_γ, S_γ) and $(\beta_\gamma, \delta_\gamma)$ are the corresponding submatrices of (X, S) and subvectors of (β,δ), and $(X_{o,\gamma}, S_{o,\gamma})$ and $(X_{c,\gamma}, S_{c,\gamma})$ the submatrices of the observed (X_o, S_o) and censored (X_c, S_c) design matrices. One then obtains posterior model probabilities

$$\begin{aligned}(5)\qquad \pi(\gamma \mid y) &= \frac{p(y \mid \gamma)\pi(\gamma)}{\sum_\gamma p(y \mid \gamma)\pi(\gamma)}\\ &= \Big(1 + \sum_{\gamma'\neq\gamma} B_{\gamma',\gamma}\frac{\pi(\gamma')}{\pi(\gamma)}\Big)^{-1},\end{aligned}$$

where $\pi(\gamma)$ is the model prior probability, $B_{\gamma',\gamma} = p(y \mid \gamma')/p(y \mid \gamma)$ the Bayes factor between (γ',γ) and

$$\begin{aligned}p(y \mid \gamma) = \int & p(y \mid \alpha_\gamma, \kappa_\gamma, \tau)\\ &\times \pi(\alpha_\gamma, \kappa_\gamma, \tau \mid \gamma)\,d\alpha_\gamma\,d\kappa_\gamma\,d\tau,\end{aligned}$$

the integrated likelihood $p(y \mid \alpha_\gamma, \kappa_\gamma, \tau)$ with respect to a prior density $\pi(\alpha_\gamma, \kappa_\gamma, \tau \mid \gamma)$. One may choose the model with highest $\pi(\gamma \mid y)$, variables with high marginal posterior probabilities $\pi(\gamma_j \neq 0 \mid y)$ and, when the interest is in prediction, use Bayesian model averaging where models are weighted according to $\pi(\gamma \mid y)$, or alternatively choosing a sparse model giving similar predictions [14]. Either way $\pi(\gamma \mid y)$ are critical for inference, hence the importance to understand their behavior.

To conclude, we comment upon a practically relevant computational issue. In additive models, it is common to either let the basis dimension r grow with n and add a regularization term (e.g., P-splines), or to learn r from the

data (e.g., knot selection). Letting r grow with n is interesting theoretically and in prediction problems where one fits a single model, but less so when one considers many models. Large r increases the computational cost (e.g., matrix determinants require $r^3/3$ operations) and is often unneeded when the goal is just to detect if a covariate has an effect. Instead one may use a moderate r, for example, misspecify the predictive-optimal model. The question is then, what answer can one hope to obtain and what are its properties. Our theory and software allow learning r among several fixed values, but in our examples a small $r = 5$ provided better inference at lower cost (particularly for small n, e.g., Figure S3, bottom).

2.3 Prior Distributions

Although our discussion applies to a wide class of priors, we present three concrete options:

$$\pi_L(\alpha_\gamma, \kappa_\gamma, \tau) = \pi(\tau) \prod_{\gamma_j \geq 1} N\big(\alpha_j; 0, g_L n/(x_j^\top x_j)\big) \times \prod_{\gamma_j = 2} N\big(\kappa_j; 0, g_S n (S_j^\top S_j)^{-1}\big)$$

$$\pi_M(\alpha_\gamma, \kappa_\gamma, \tau) = \pi(\tau) \prod_{\gamma_j \geq 1} \frac{\alpha_j^2}{g_M} N(\alpha_j; 0, g_M) \times \prod_{\gamma_j = 2} N\big(\kappa_j; 0, g_S n (S_j^\top S_j)^{-1}\big)$$

$$\pi_E(\alpha_\gamma, \kappa_\gamma, \tau) = \pi(\tau) \prod_{\gamma_j \geq 1} e^{\sqrt{2} - g_E/\alpha_j^2} N(\alpha_j; 0, g_E) \times \prod_{\gamma_j = 2} N\big(\kappa_j; 0, g_S n (S_j^\top S_j)^{-1}\big),$$

where $\pi(\tau) = 2\tau^{-3}\mathrm{IG}(\tau^{-2}; a_\tau/2, b_\tau/2)$, and IG denotes the inverse gamma density, and $g_L, g_S, g_M, g_E, a_\gamma, b_\tau \in \mathbb{R}_+$ are given dispersion parameters, for which we propose default values in Section 2.4.

These choices include a standard Normal prior and two variations of nonlocal priors. The use of nonlocal priors can be argued from a foundational viewpoint, where one wishes to assign prior beliefs that are coherent with the parameters assumed nonzero by a given model [23]. For our purposes, however, their main role is that they lead to faster Bayes factor rates to discard spurious parameters. See [41] for further discussion. We refer to π_L as group-Zellner prior. It is a product of Zellner priors across groups of linear and nonlinear terms for each covariate. This prior is local, that is, it assigns nonzero density to α_γ having zeroes. The Zellner structure is chosen for simplicity, our theory can be easily extended to other local priors, provided they are continuous and positive at the asymptotically optimal parameter values (Section 3, [23]). The priors π_M and π_E are nonlocal with respect to α_γ, the so-called product MOM and eMOM priors introduced in [24, 42], and a group-Zellner prior on κ_γ.

Regarding the prior on the models $\pi(\gamma)$, we consider joint group inclusion of nonlinear coefficients δ_j and the hierarchical restriction that their inclusion requires that of the corresponding linear coefficient β_j. Letting $\pi(\gamma)$ depend only on the number of nonzero parameters in $(\beta_\gamma, \delta_\gamma)$, as customarily done when only linear effects are considered, would ignore such structure and hence be inadequate. Instead, we let $\pi(\gamma)$ depend on the number of variables having linear and nonlinear effects, (p_γ, s_γ). By default, we consider independent Beta-Binomial priors [45]

$$\pi(\gamma) = \frac{1}{C}\mathrm{BetaBin}(p_\gamma; p, a_1, b_1)\binom{p}{p_\gamma}^{-1} \times \mathrm{BetaBin}(s_\gamma; s, a_2, b_2)\binom{s}{s_\gamma}^{-1}, \tag{6}$$

where $\mathrm{BetaBin}(z; p, a, b)$ is the probability of z successes under a Beta-Binomial distribution with p trials and parameters (a, b) and C a normalizing constant that does not need to be computed explicitly. Any model such that the number of parameters is $p_\gamma + r s_\gamma > n$ is assigned $\pi(\gamma) = 0$, as it would result in data interpolation. By default, we let $a_1 = b_1 = a_2 = b_2 = 1$ akin to [45], for example, in the $p = 1$ case these give $\pi(\gamma_1 = 0) = \pi(\gamma_1 = 1) = \pi(\gamma_1 = 2) = 1/3$. As alternatives to (6), one can also consider Binomial priors where $\mathrm{BetaBin}(z; p, a_j, b_j)$ is replaced by $\mathrm{Bin}(z; p, a_j)$ for a given success probability $a_j \in [0, 1]$ and Complexity priors [5] where it is replaced by $1/p^{a_j z}$ for some constant $a_j > 0$. These two alternatives are implemented in our software and covered by our theory in Section 3, but for simplicity our examples focus on (6).

2.4 Prior Elicitation

The prior dispersion parameters (g_L, g_M, g_E, g_S) are important for variable selection. For instance, setting large dispersions helps induces sparsity, particularly when they are allowed to grow with the sample size n [32]. However, such large values also reduce power (see [37] and our Propositions 3, 4 and S5) and are harder to justify from the point of view that the expected effect sizes a priori should not depend on n. We briefly discuss default values that do not depend on n, and refer the reader to Section S3 for details.

Specifying prior parameters provides an opportunity to define what effects are practically relevant. Importantly, in what follows we assume that continuous covariates were standardized to unit variance, else the parameter interpretation and default values change. Basic considerations

give a fairly narrow range of values that we deem reasonable in applications. For example, in AFT and Cox models $e^{|\beta_j|}$ define the effect size, when these are say $< 15\%$ (i.e., $e^{|\beta_j|} < 1.15$) they are typically practically irrelevant. Based on these considerations, our recommended defaults for AFT and Cox models are $g_M = 0.192$, $g_E = 0.091$, $g_L = 1$, $g_S = 1/r$ and $a_\tau = b_\tau = 3$, whereas for probit regression they are $g_M = 0.139$ and $g_E = 0.048$. One should not take these defaults at their exact value, rather as defining a range of reasonable values. These ranges are discussed in Section S3. In our examples, results were robust to the prior dispersions, provided they stay within our recommended range.

We remark that if one were to change the prior dispersion arbitrarily then results would be affected, in a similar manner to how regularization parameters affect penalized likelihood results. However, in our view the prior beliefs implied by arbitrary prior dispersions would be unreasonable in most applications. We also note that there is a wide objective Bayes literature on using the data to set the prior parameters; see [7] for an excellent review. We do not argue against such strategies, but we focus on our defaults as a simple strategy that attains a fairly competitive performance in practice.

3. THEORY

This section describes the asymptotic solution returned by Bayesian model selection, when the observed data $(o_i, c_i, z_i) \sim F_0$ are independent realizations from some F_0, where $z_i \in \mathbb{R}^{p(r+1)+q}$ for $q \geq 0$ contains the observed covariates $(x_i, s_i) \in \mathbb{R}^{p(r+1)}$, and potentially also q additional columns. These columns may contain covariates that were not recorded but are truly relevant for the outcome or the censoring, or nonlinear effects and interactions missed by (x_i, s_i). We do not assume F_0 to be parametric, rather it can be quite general, and the whole model structure assumed by the analyst (e.g., accelerated times, proportional hazards) may be wrong.

Section 3.1 shows that when one assumes the Normal AFT model (1) but truly $(o_i, c_i, z_i) \sim F_0$, the maximum likelihood estimator under each model γ converges to an optimal $(\alpha_\gamma^*, \kappa_\gamma^*, \tau_\gamma^*)$ and is asymptotically normally distributed. See [17] and [18] for related asymptotic results, and Section S8 for analogous results for the Laplace AFT model. Section 3.2 shows that Bayesian model selection in the AFT model asymptotically returns the smallest γ^* such that all effects in $(\alpha_\gamma^*, \kappa_\gamma^*)$ are nonzero. Equivalently, γ^* is defined by the zeroes in (α^*, κ^*), the optimal value under the full model including all parameters. Section 3.3 gives analogous results for Cox models. These results are extended to probit models in Section S9, and in Section S10 to more general concave log-likelihood models. It is possible to derive similar results beyond the concave case, however, this class encompasses all the models we consider here and allows simplifying the proofs and technical conditions.

Throughout we help interpret the solution and certain Bayes factors properties. Of particular relevance, Section 3.1 discusses that the asymptotic solution γ^* excludes covariates that do not help predict the outcome nor the censoring times, and offers some examples. Section 3.2 comments on potential advantages of using low-dimensional basis and nonlinear decompositions to detect covariate effects.

3.1 Asymptotic Solution in AFT Models

As the sample size grows, Bayesian model selection recovers a model γ^* that excludes parameters that are asymptotically estimated to be zero. Under mild regularity conditions, this limiting parameter is the value maximizing the expected log-likelihood under F_0. We start by defining the expected log-likelihood, then state the limiting result, and finally interpret its meaning and implications for model selection.

Let $\eta_\gamma = (\alpha_\gamma, \kappa_\gamma, \tau) \in \Gamma_\gamma$ be the vector with $p_\gamma + rs_\gamma$ regression parameters under a given model γ (Section 2.2) plus the error variance, where $\Gamma_\gamma = \mathbb{R}^{p_\gamma + rs_\gamma} \times \mathbb{R}^+$ is the corresponding parameter space. Let

$$\begin{aligned} m(\eta_\gamma) = (1-u_1)\big[\log\Phi\big(x_1^\top\alpha_\gamma + s_1^\top\kappa_\gamma - \tau\log(c_1)\big)\big] \\ + u_1\bigg[\log(\tau) - \frac{1}{2}\log(2\pi) \\ - \frac{1}{2}\big(\tau\log(o_1) - x_1^\top\alpha_\gamma - s_1^\top\kappa_\gamma\big)^2\bigg], \end{aligned}$$

the contribution of one observation to the log-likelihood (2), and

$$\begin{aligned} M(\eta_\gamma) &= \mathbb{E}_{F_0}\big(m(\eta_\gamma)\big) \\ &= P_{F_0}(u_1 = 0) \\ &\quad \times \mathbb{E}_{F_0}\big[\log\Phi\big(x_{1\gamma}^\top\alpha_\gamma + s_{1\gamma}^\top\kappa_\gamma \\ &\quad - \tau\log(c_1)\big) \mid u_1 = 0\big] \\ &\quad + P_{F_0}(u_1 = 1)\bigg(\log(\tau) - \frac{1}{2}\log(2\pi) \\ &\quad - \frac{1}{2}\mathbb{E}_{F_0}\big[\big(\tau\log(o_1) - x_{1\gamma}^\top\alpha_\gamma \\ &\quad - s_{1\gamma}^\top\kappa_\gamma\big)^2 \mid u_1 = 1\big]\bigg) \end{aligned} \tag{7}$$

its expectation under the data-generating F_0. Under minimal conditions, $M(\eta_\gamma)$ has a unique maximizer, denoted by $\eta_\gamma^* = (\alpha_\gamma^*, \kappa_\gamma^*, \tau_\gamma^*)$. Below we focus our interpretation on viewing (7) as the expectation of a likelihood-associated reward, and η_γ^* as the associated minimizer, but η_γ^* can also be viewed as minimizing the Kullback–Leibler divergence to $F_0(y, u)$ (also called generalized Kullback–Leibler divergence; see [17]).

Proposition 1 proves that the maximum likelihood estimator $\widehat{\eta}_\gamma$ converges to η_γ^*, and Proposition 2 its asymptotic normality with a sandwich covariance that is standard in misspecified models, and corresponds to the smallest possible covariance for unbiased estimators under model misspecification. Such variance alteration does not affect consistency but can alter finite n false positives and asymptotic power (see Section 3.2). See also Propositions S1–S2 for analogous results on the AFT model with Laplace errors. Mild technical conditions, denoted A1–A5, that suffice for the proposition to hold are discussed in S7. We remark that A3 assumes the existence and finiteness of η_γ^* and $\widehat{\eta}_\gamma$ (the latter for large enough n), which implies that these optima cannot occur at the boundary of Γ_γ and must be unique (by concavity). For example, this rules out situations where η_γ^* contains infinite regression parameters or variance, or zero variance, which we view as pathological cases that we exclude from consideration. Similar assumptions were made by [2] (Assumption M in Section 3; see also references therein), although those authors allowed for the maximum to occur on the boundary of Γ_γ.

PROPOSITION 1. *Assume A1–A3. Then* $\eta_\gamma^* = \operatorname{argmax}_{\Gamma_\gamma} M(\eta_\gamma)$ *is unique and* $\widehat{\eta}_\gamma \xrightarrow{P} \eta_\gamma^*$ *as* $n \to \infty$.

PROPOSITION 2. *Assume A1–A5. Then*

$$\sqrt{n}(\widehat{\eta}_\gamma - \eta_\gamma^*) \xrightarrow{D} N\big(0, V_{\eta_\gamma^*}^{-1}\mathbb{E}_{F_0}\big[\nabla m(\eta_\gamma^*)\nabla m(\eta_\gamma^*)^\top\big]V_{\eta_\gamma^*}^{-1}\big),$$

where $V_{\eta_\gamma^*}$ *is the Hessian matrix of* $M(\eta_\gamma)$ *evaluated at* η_γ^*, *and* $m(\eta_\gamma^*) = \log p(y_1 \mid \eta_\gamma^*)$.

Proposition 1 has important implications for model selection. Let (α^*, κ^*) be the optimal parameter under the full model that includes all linear and nonlinear terms. Asymptotically, one obtains the model γ^* of smallest dimension maximizing (7) (see Section 3.2), which is defined by zeroes in (α^*, κ^*). Specifically, $\gamma_j^* = 0$ if both linear and nonlinear coefficients (α_j^*, κ_j^*) are zero, $\gamma_j^* = 1$ if $\alpha_j^* \neq 0$ and $\kappa_j^* = 0$, and $\gamma_j^* = 2$ if $\kappa_j^* \neq 0$.

To interpret this asymptotic solution, we turn attention to (7). If a covariate does not contribute to improving neither of the two terms in (7), then its corresponding entry in (α^*, κ^*) is zero. The first term is the expected log-probability, as predicted by the model, that the individual is censored at the observed $\log(c_1)$ (conditional on being censored). Therefore, any covariate that helps the model predict more accurately the occurrence of censoring events contributes to this first term. The second term is the mean squared error in predicting the observed time $\log(o_1)$, conditional on the time being uncensored. Expression (7) is an average of these two components weighted by the true censoring probability $P_{F_0}(u_1 = 0)$, and averaged across covariate values under F_0. Hence, γ^* drops covariates that do not predict survival neither censoring times, but may include those that, even if truly unrelated to survival, help explain the censoring. This interpretation extends to working models other than the Normal AFT. For any other fully specified model, the first term in (7) features the model log-predicted probability of censoring, and the second term the usual log-likelihood for uncensored data. For example, under a AFT model with Laplace errors the asymptotic solution is defined by the mean absolute error and the Laplace survival function (see Section S8).

We present some simple examples to illustrate our discussion.

EXAMPLE. Suppose that under F_0, truly $\log o_i \mid c_i \sim N(x_{i1} + \theta \log c_i, \sigma^2)$ and $\log c_i \sim N(x_{i2}, \sigma^2)$. The analyst adopts the model $\log o_i \sim N(\beta_1 x_{i1} + \beta_2 x_{i2}, 1/\tau^2)$, which, as discussed, assumes noninformative censoring. If $\theta = 0$, the censoring under F_0 is noninformative, and then $\alpha_2^* = \beta_2^* = 0$, hence x_{i2} is discarded asymptotically.

However, if $\theta \neq 0$ then truly $\log o_i = x_{i1} + \theta x_{i2} + \epsilon_i$, where $\epsilon_i \sim N(0, (1 + \theta^2)\sigma^2)$. Plugging this expression into (7), it is easy to show that then $\alpha_2^* \neq 0$. That is, the presence of informative censoring causes x_{i2} to be asymptotically selected.

EXAMPLE. Suppose that there is a fixed administrative censoring at $\log c_i = a$ for all individuals (so it is truly noninformative under F_0), a single covariate $x_i \in \mathbb{R}$, and that the analyst adopts the model $\log o_i \sim N(\beta_1 + \beta_2 x_i, 1/\tau^2)$. Suppose that x_i truly has an effect on the outcome under F_0, but that said effect only occurs at a time $b > a$. Then the effect cannot be detected from the observed data, since all individuals are censored at a. The issue is that the covariate has an effect that deviates from the assumed AFT structure. For example, suppose that under F_0,

$$\log o_i = z_i + \theta x_i \mathrm{I}(z_i > b),$$

where $x_i \in \{0, 1\}$ indicates that individual i received a treatment, $z_i \sim N(0, 1)$ is the survival time for untreated individuals, and $\theta > 0$ quantifies the treatment effect.

Here the effect is only present among individuals that live longer than b and, since censoring occurs before b, for all uncensored individuals one observes $\log o_i = z_i$. Plugging this expression and $\log c_i = a$ into (7), and noting that the conditioning on u_1 can be removed from the expectations, one can show that $\alpha_2^* = \beta_2^* = 0$. This is an extreme example where one cannot detect an effect that strongly deviates from the assumed mean structure, even though the censoring is noninformative. One could conceive related examples where a covariate has a time-varying effect that is first positive and then negative, before administrative censoring occurs, so that the average effect is near-zero.

EXAMPLE. Suppose that a potentially informative censoring occurs early, so that $P_{F_0}(u_1 = 0) \approx 1$. Then (7) under the full model is approximately equal to

$$\mathbb{E}_{F_0}\big[\log \Phi\big(x_1^\top \alpha + s_1^\top \kappa - \tau \log(c_1)\big)\big].$$

As discussed, this term is the log-probability that the outcome occurs after the observed censoring time, as predicted by the Normal AFT model. Hence, (α^*, κ^*) are essentially chosen to predict censoring times. If the censoring is informative and depends on a set of covariates, then (α^*, κ^*) will in general assign nonzero coefficients to these covariates, which will be asymptotically selected. A similar argument can be made for late censoring where $P_{F_0}(u_1 = 1) \approx 1$, then (α^*, κ^*) is approximately the usual (population) least-squares solution. If the outcome depends on the censoring, which in turn depends on a set of covariates, then least-squares will assign a nonzero coefficient to the latter.

3.2 Bayes Factor Rates for Misspecified AFT Models

This section proves that the posterior probability of the optimal model γ^* converges to 1, under mild conditions. Recall that the posterior probability of γ^* is

$$\pi(\gamma^* \mid y) = \frac{p(y \mid \gamma^*)\pi(\gamma^*)}{\sum_\gamma p(y \mid \gamma)\pi(\gamma)} = \left(1 + \sum_{\gamma \neq \gamma^*} B_{\gamma,\gamma^*} \frac{\pi(\gamma)}{\pi(\gamma^*)}\right)^{-1}.$$

Proposition 3 gives the rate at which each B_{γ,γ^*} converges to 0 (in probability), when ones assumes a potentially misspecified AFT model. Provided that each $B_{\gamma,\gamma^*}\pi(\gamma)/\pi(\gamma^*)$ converges to 0 (this follows immediately in the standard case where prior model probabilities are bounded, e.g.) it follows that $\pi(\gamma^* \mid y) \xrightarrow{P} 1$. This implies that the highest posterior probability model consistently selects γ^*, and that including covariates with marginal posterior probability $\pi(\gamma_j^* \mid y) > t$, for any fixed threshold t, also leads to consistent selection.

Proposition 3 clarifies the role of censoring and misspecification. The result is stated for Laplace approximations to Bayes factors, a computationally convenient alternative to obtaining exact marginal likelihoods, but in our setting both are asymptotically equivalent (Proposition S6). Specifically, we consider

$$B_{\gamma,\gamma^*} = \frac{\widehat{p}(y \mid \gamma)}{\widehat{p}(y \mid \gamma^*)}, \tag{8}$$

where $\widehat{p}(y \mid \gamma)$ is obtained via a Laplace approximation:

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

where $\tilde{\eta}_\gamma = \arg\max_{\eta_\gamma} \ell(\eta_\gamma) + \log \pi(\eta_\gamma)$ is the maximum a posteriori under prior $\pi(\eta_\gamma)$. See Section S4 for details on computing this approximation.

Proposition 3 treats separately overfitted models (containing γ^*) and nonoverfitted models (not containing γ^*). Overfitted models contain all truly relevant plus a few spurious parameters, a situation where the challenge is to enforce sparsity. Nonoverfitted models are missing some truly relevant parameters, there the challenge is also to have high power to detect the missing signal. By truly relevant we mean improving $M(\eta_\gamma^*)$, that is, the prediction of either observed or censored times; see Section 3.1. Recall that $d_\gamma = \dim(\eta_\gamma) = p_\gamma + rs_\gamma + 1$. Intuitively the proof of Proposition 3 is based on establishing the asymptotic distribution of the likelihood-ratio test statistic $2[\ell(\tilde{\eta}_\gamma) - \ell(\tilde{\eta}_{\gamma^*})]$, which is bounded by central chi-squares in the overfitted case and noncentral chi-squares in the nonoverfitted case, and then finding an asymptotic approximation to the other quantities featuring in $\widehat{p}(y \mid \gamma)$.

PROPOSITION 3. *Let B_{γ,γ^*} be the Bayes factor in (8) under either π_L, π_M or π_E, where γ^* is the AFT model with smallest d_{γ^*} minimizing (7), and $\gamma \neq \gamma^*$ another AFT model. Assume that both γ^* and γ satisfy Conditions A1–A5. Suppose that (g_M, g_E, g_L, g_S) are nondecreasing in n.*

(i) *Overfitted models. If $\gamma^* \subset \gamma$, then*

$$\log B_{\gamma\gamma^*} = \log(a_n) + \frac{r}{2}(s_{\gamma^*} - s_\gamma)\log(ng_S) + \mathcal{O}_{\mathrm{p}}(1),$$

where $a_n = (ng_L)^{\frac{p_{\gamma^*} - p_\gamma}{2}}$ under π_L, $a_n = (n \times g_M)^{3(p_{\gamma^*} - p_\gamma)/2}$ under π_M, and $a_n = (ng_E \times e^{2g_E\sqrt{n}})^{(p_{\gamma^*} - p_\gamma)/2}$ under π_E.

(ii) *Nonoverfitted models. If $\gamma^* \not\subset \gamma$, then*

$$\log(B_{\gamma\gamma^*}) = -n\big[M(\eta_{\gamma^*}^*) - M(\eta_\gamma^*)\big] + \log(b_n) + \frac{r}{2}(s_{\gamma^*} - s_\gamma)\log(ng_S) + \mathcal{O}_{\mathrm{p}}(1)$$

where $b_n = (ng_L)^{\frac{p_{\gamma^*} - p_\gamma}{2}}$ under π_L, $b_n = (n \times g_M^3)^{(p_{\gamma^*} - p_\gamma)/2}$ under π_M, and $b_n = (g_E n)^{p_{\gamma^*} - p_\gamma} \times e^{-g_E c}$ under π_E, for finite $c \in \mathbb{R}$.

By Proposition 3(i) the rates to discard overfitted models are unaffected by misspecification and censoring (but certain constants can affect finite n behaviour; see the proof). These sparsity rates are improved by nonlocal priors and by setting large prior dispersions (g_L, g_M, g_E, g_S), extending previous results [24, 32, 39, 41] to misspecified survival models. By Proposition 3(ii) the rate to detect nonspurious effects is exponential in n with a coefficient $M(\eta_{\gamma^*}^*) - M(\eta_\gamma^*) > 0$ that measures the drop of predictive ability in γ relative to γ^*, and is hence affected by misspecification and censoring. Recall that

predictive ability can be understood as a weighted average of forecasting the outcome to occur after the censoring time (for censored individuals) and the actual outcome time (for uncensored individuals).

When one misspecifies the model family, $M(\eta^*_{\gamma^*}) - M(\eta^*_\gamma)$ is driven by the projection of F_0 onto the assumed family. Interpreting the geometry of such projections is beyond our scope, but intuitively projections usually reduce distances and hence make $M(\eta^*_{\gamma^*}) - M(\eta^*_\gamma)$ smaller than if one were to assume the correct model class. By Part (ii), this would decrease the power to detect nonzero effects in η^*_γ.

To facilitate interpretation suppose there is no censoring. Then simple algebra shows that $M(\eta^*_{\gamma^*}) - M(\eta^*_\gamma) = \mathbb{E}_{F_0}[\log(\tau^*_{\gamma^*}/\tau^*_\gamma)]$, which measures the difference in mean squared prediction errors from using model γ instead of the optimal γ^* (given by $1/(\tau^*_\gamma)^2$ and $1/(\tau^*_{\gamma^*})^2$, respectively). For instance, omitting covariates increases $\tau^*_{\gamma^*}/\tau^*_\gamma$, causing an exponential drop in power; see our examples in Sections 5.1-5.2 for an illustration.

Proposition 3 also highlights trade-offs in modeling nonlinear covariate effects. Including a truly active nonlinear term is rewarded by an improved model fit $M(\eta^*_{\gamma^*}) - M(\eta^*_\gamma)$, but runs into an $r\log(ng_S)$ penalty. In contrast, including a linear effect leads to a smaller improvement in fit, but also incurs a smaller $\log(ng_S)$ penalty. Hence, decomposing effects into a linear and nonlinear components can improve power.

A similar observation illustrates that for model selection purposes, the advantages of using fully nonparametric effects over a finite-dimensional basis may be small. Suppose one replaced the basis dimension r by a larger r^* maximizing $M(\eta^*_{\gamma^*}) - M(\eta^*_\gamma)$. For m-degree splines with equi-spaced knots and sufficiently smooth $M()$ the improvement in $M(\eta^*_{\gamma^*}) - M(\eta^*_\gamma)$ associated to increasing r to r^* is at most of order $1/r^m$ [36]. For said increase to offset the complexity penalty it needs to hold that $r^{m+1}(r^*-r)/2$ is of a smaller order than $n/\log(ng_S)$. Hence, by letting $r^{m+1}r^*$ grow sub-linearly with n could improve power relative to r. However, for even moderate r and cubic splines ($m = 3$) the required $n > r^*r^4$ can be impractically large; see, for example, the examples in Section 5.1 with $r \in \{5, 10, 15\}$. Further, the computational cost of using a large r^* for each considered model γ is impractical when one wishes to consider many models.

In summary, using a small basis dimension r (e.g., $r = 5$, in our examples) within the nonlinear effect decomposition in Section 2.1 may be practically preferable to a nonparametric basis where r grows with n, for the purpose of detecting the effect.

3.3 Bayes Factor Rates for Misspecified Additive Cox Models

Our Bayes factor results under misspecified Cox models are similar to Section 3.2, but here the optimal model γ^* is defined by zeroes in the parameter $\eta^* = (\beta^*, \delta^*)$ maximizing the expected partial likelihood (3) under F_0; see S7.11 for its expression and some discussion. The interpretation of η^* is also analogous, though here (3) rewards predicting a higher risk for individuals who experienced the event (uncensored) than for other individuals at risk. An alternative interpretation is possible by noting that (3) can be approximated by a Poisson regression log-likelihood [27], where one models the mean number of uncensored events in infinitesimal intervals. Intuitively, any covariate that helps predict this mean, which depends on the distribution of the censoring and survival times, is asymptotically selected. Covariates that are unrelated both to survival and censoring are hence discarded.

We consider Bayes factors obtained by a Laplace approximation to the integrated partial likelihood

$$p(y \mid \gamma) = \int \exp\{\ell_p(\beta_\gamma, \delta_\gamma)\} \times \pi(\beta_\gamma, \delta_\gamma \mid \gamma)\, d\beta_\gamma\, d\delta_\gamma \tag{9}$$

these can be viewed as the integrated likelihood under a limiting noninformative nonparametric Gamma process prior on h_0; see [28] and [33] for a discussion. We obtain Bayes factor rates analogous to Section 3.2, the proof builds upon [54] and [30] who proved that $\bar\eta_\gamma = (\bar\beta_\gamma, \bar\delta_\gamma)$ maximizing (3) are consistent and asymptotically normal under misspecification, under Conditions B1–B4 listed in Section S7.4.

PROPOSITION 4. *Let B_{γ,γ^*} be the Bayes factor based on (9) under π_L, π_M or π_E, γ^* the Cox model with smallest d_{γ^*} minimizing the expected log partial likelihood M_p in (S7.10) and $\gamma \neq \gamma^*$ another Cox model. Assume that (γ^*, γ) satisfy Conditions B1–B4, and that (g_M, g_E, g_L, g_S) are nondecreasing in n.*

(i) *Let a_n be as in Proposition 3. If $\gamma^* \subset \gamma$, then*

$$\log B_{\gamma\gamma^*} = \log(a_n) + \frac{r}{2}(s_{\gamma^*} - s_\gamma)\log(ng_S) + \mathcal{O}_p(1),$$

(ii) *Let b_n be as in Proposition 3. If $\gamma^* \not\subset \gamma$, then*

$$\log(B_{\gamma\gamma^*}) = -n\big[M_p(\eta^*_{\gamma^*}) - M_p(\eta^*_\gamma)\big] + \log(b_n) + \frac{r}{2}(s_{\gamma^*} - s_\gamma)\log(ng_S) + \mathcal{O}_p(1).$$

That is, the Bayes factors under an assumed Cox model have similar asymptotic behavior as under an assumed AFT model, hence the conclusions stated in Section 3.2 also apply to the Cox model.

4. COMPUTATION

The two main computational challenges are exploring the model space $\gamma \in \{0, 1, 2\}^p$, and approximating the integrated likelihood $p(y \mid \gamma)$ in (5) for each model. We

first discuss relative advantages of the Normal AFT and Cox models for computing $p(y \mid \gamma)$, and how they relate to the amount of censored data in Section 4.1. We also discuss an approximation to the Normal log-distribution function derivatives that dramatically speeds up computation for the AFT and probit models. Section 4.2 discusses the model search, when one cannot enumerate all 3^p models.

4.1 Within-Model Calculations

When the log-likelihood is concave (or locally concave around η_γ^*, as in asymptotically Normal models), Laplace approximations to $p(y \mid \gamma)$ are one of the fastest and more accurate methods available. A practical limitation is that, when one wishes to consider many models or the sample size is large, solving the required optimization problems can still be cumbersome. This cost can be significantly ameliorated by combining convex optimization algorithms that use warm initializations; see Section S4. See also [38] for an approach based on approximate Laplace approximations that bypasses the optimization exercise altogether.

Within survival analysis, an advantage of exponential-family AFT models is admitting sufficient statistics for the uncensored part of the likelihood, for example, $(y_o^\top y_o, X_o^\top y, X_o^\top X_o)$ for (2). These can be computed upfront in $n_o(1+p+p(p+1)/2)$ operations and re-used whenever a new model γ is considered at no extra cost, but for large p such pre-computation has significant cost and memory requirements. Since one typically visits only a small subset of models, many elements in $X_o^\top X_o$ are never used and it would be wasteful to compute them all upfront. It is more convenient to compute the entries in $X_o^\top X_o$ when first required by any given γ and storing them for later use. Our software follows this strategy by using sparse matrices in the `C++ Armadillo` library [43].

Given these sufficient statistics the log-likelihood in (2) requires $\min\{nd_\gamma, (n_c+1)d_\gamma + d_\gamma(d_\gamma+1)/2\}$ operations, and each entry in its gradient and hessian require n_c+1 further operations. In contrast the Cox model's partial likelihood has a minimum cost of $n_o d_\gamma + n_o(n_o-1)/2$ operations when censored times precede all observed times ($\max c_i < \min o_i$), and a maximum cost $nd_\gamma + [n(n+1) - n_c(n_c-1)]/2$ when observed times precede all censored times. That is, the AFT likelihood has a significantly lower cost than the Cox model when $n_c < n_o$ (moderate censoring) or $n > d_\gamma$ (sparse settings).

A caveat of the Normal AFT model, however, is requiring the extensive evaluation of the log-cumulative distribution $\log\Phi$ and its derivatives. Each likelihood evaluation requires n_c terms featuring Φ and, although these terms can be re-used when computing $r(z) = \phi(z)/\Phi(z)$ and $D(z) = r(-z)^2 - zr(-z)$ in the gradient and hessian, evaluating $\Phi(z)$ is costly. Briefly, the problem of approximating the inverse Mill's ratio $r(z)$ has been well studied [12]. There are many algorithms to approximate $\Phi(z)$, but $r(z)$ is harder, for example, Expression 26.2.16 in [1] (page 932) has maximum absolute error $< 7.5\times10^{-8}$ for $\Phi(z)$ but unbounded absolute error for $r(z)$ as $z\to-\infty$. By combining existing proposals we built a fast approximation that guarantees the small relative errors. One may combine the Taylor series and asymptotic expansions in [1] (page 932, Expressions 26.2.16 and 26.2.12) for $\Phi(z)$ with an optimized Laplace continued fraction in [29] (Expression (5.3)) for $r(z)$ as $z\to-\infty$. The resulting $\widehat{r}(z)$ has maximum absolute and relative errors < 0.000185 and < 0.000102 respectively, and for $\widehat{D}(z) = \widehat{r}(-z)^2 - z\widehat{r}(-z)$ they are < 0.000424 and < 0.000505. See Section S5 for further details. As an empirical check, the posterior model probabilities obtained in Section 5.3 when replacing $(r(z), D(z))$ by $(\widehat{r}(z), \widehat{D}(z))$ remained identical to the third decimal place.

This approximation also facilitates evaluating the log-likelihood and derivatives for probit and other models involving $\log\Phi$, and may have some independent interest. Using this approximation and the warm initializations in Section S4 is practically meaningful, for the TGFB data (Section 5.3, 868 parameters) they reduced the cost of 1,000 Gibbs iterations from >4 hours to 38 seconds.

4.2 Model Exploration

Recent advances in Markov chain Monte Carlo provide model exploration strategies that perform fairly well in practice; see [59] for a tempering approach that is particularly helpful when there are multimodalities in $p(\gamma \mid y)$, or [13] for adaptive methods that reduce the effort in exploring low posterior probability models. Further, as n grows and posterior probabilities concentrate on a single model, it is possible to prove quick convergence [57]. Intuitively, if $p(\gamma^* \mid y) \approx 1$ and the chain converges quickly, there is high probability that γ^* will be visited after a few iterations. Most iterations are spent on models with high $\pi(\gamma \mid y)$ which, from Proposition 3, are models with dimension close to d_{γ^*}. The main burden arises from obtaining $p(y \mid \gamma)$, which only needs to be computed the first time that γ is visited and can be stored for subsequent iterations. Hence, if d_{γ^*} is not too large (sparse data-generating truths) or $\pi(\gamma \mid y)$ is concentrated on relatively few models, the cost is manageable.

Here for simplicity we describe Algorithm 1, a Gibbs algorithm that builds upon earlier proposals [24, 39], with the novelty that it adds a latent augmentation to enforce hierarchical restrictions (nonlinear terms in S are only added if the corresponding linear term in X is in the model) in a computationally efficient manner. The algorithm obtains B samples $\gamma^{(1)}, \ldots, \gamma^{(B)}$ from $\pi(\gamma \mid y)$. It is not a naive Gibbs algorithm that sequentially

Algorithm 1 Augmented-space Gibbs sampling

1: Set $b = 0$, $\tilde{\gamma}^{(0)} = (0, \ldots, 0)$.

2: For $j = 1, \ldots, 2p$, update $\tilde{\gamma}_j^{(0)} = \arg\max_k p(\tilde{\gamma}_j = k \mid y, \tilde{\gamma}_{-j}^{(b)})$. If an update was made across $j = 1, \ldots, 2p$ go back to Step 2, else set $\gamma_j^{(0)} = \max\{\gamma_j^{(0)}, \gamma_{j+p}^{(0)}\}$ for $j = 1, \ldots, p$ and go to Step 3.

3: Set $b = b + 1$. For $j = 1, \ldots, p$ set $\tilde{\gamma}_j^{(b)} = 1$ with probability

$$P(\tilde{\gamma}_j = 1 \mid y, \tilde{\gamma}_{-j}^{(b)}) = \begin{cases} 1, & \text{if } \tilde{\gamma}_{j+p} = 1, \\ \dfrac{p(y \mid \tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)}) p(\tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)})}{p(y \mid \tilde{\gamma}_j = 0, \tilde{\gamma}_{-j}^{(b)}) p(\tilde{\gamma}_j = 0, \tilde{\gamma}_{-j}^{(b)}) + p(y \mid \tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)}) p(\tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)})}, & \text{if } \tilde{\gamma}_{j+p} = 0, \end{cases}$$

and otherwise set $\tilde{\gamma}_j^{(b)} = 0$.

4: For $j = p + 1, \ldots, 2p$ set $\tilde{\gamma}_j^{(b)} = 1$ with probability

$$P(\tilde{\gamma}_j = 1 \mid y, \tilde{\gamma}_{-j}^{(b)}) = \begin{cases} 0, & \text{if } \tilde{\gamma}_{j+p} = 0, \\ \dfrac{p(y \mid \tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)}) p(\tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)})}{p(y \mid \tilde{\gamma}_j = 0, \tilde{\gamma}_{-j}^{(b)}) p(\tilde{\gamma}_j = 0, \tilde{\gamma}_{-j}^{(b)}) + p(y \mid \tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)}) p(\tilde{\gamma}_j = 1, \tilde{\gamma}_{-j}^{(b)})}, & \text{if } \tilde{\gamma}_{j+p} = 1, \end{cases}$$

and otherwise set $\tilde{\gamma}_j^{(b)} = 0$. If $b = B$ stop, else go back to Step 3.

samples p trinary indicators, that is, sets $\gamma_j^{(b)} = k$ with probability $\pi(\gamma_j = k \mid y, \gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$ for $k \in \{0, 1, 2\}$. Instead, it is more convenient to run an augmented-space Gibbs on $2p$ binary indicators. Specifically let $\tilde{\gamma}_j = \mathrm{I}(\gamma_j = 1)$ for $j = 1, \ldots, p$ denote that covariate j only has a linear effect, and $\tilde{\gamma}_j = \mathrm{I}(\gamma_{j-p} = 2)$ for $j = p + 1, \ldots, 2p$ a nonlinear effect. Algorithm 1 samples $\tilde{\gamma}_j$ individually but prevents $(\tilde{\gamma}_j, \tilde{\gamma}_{j+p}) = (0, 1)$, that is, enforces that having a nonlinear effect when $\beta_j = 0$ has zero posterior probability. The greedy initialization of $\tilde{\gamma}^{(0)}$ is analogous to that in [24] and to the heuristic optimization in [35].

We remark that Algorithm 1 may suffer from worse mixing than naive Gibbs sampling of $\gamma_j \in \{0, 1, 2\}$, but is advantageous in sparse settings. If covariate j has a small posterior probability $\pi(\gamma_j \neq 0 \mid y)$ then $\pi(\tilde{\gamma}_j = 1 \mid y)$ is small and in most iterations $\tilde{\gamma}_{j+p}$ is set to zero without the need to perform any calculation. In contrast when sampling $\gamma_j \in \{0, 1, 2\}$ one must obtain the integrated likelihood for $\gamma_j = 2$, which can be costly due to adding the r extra parameters needed to capture the nonlinear effect. As an example, in Section 5.3 sampling $\gamma_j \in \{0, 1, 2\}$ took over 5 times longer to run than Algorithm 1, but provided the same effective sample size up to 2 decimal places.

5. EMPIRICAL RESULTS

We illustrate via examples the effect of censoring, misspecification and the use of nonlinear effect decompositions on model selection. Section 5.1 considers a simple simulation study with $p = 2$ variables, which Section 5.2 extends to $p = 50$. We consider different data-generating truths where the covariates have a monotone or nonmonotone effect, and where the truth follows an AFT, proportional hazards, or generalized hazards structure. In Section 5.3, we analyze the effect of gene TGFB on colon cancer. Given that the data-generating truth is unknown, in Section 5.4 we study the number of false positives via a permutation exercise. See also Section S11.3, where we analyze the effect of the estrogen receptor on breast cancer survival.

We consider five model selection methods combining the AFT and Cox models with local and nonlocal priors and with LASSO. For all Bayesian methods we took the highest posterior probability model $\widehat{\gamma} = \arg\max \pi(\gamma \mid y)$ as the selected model. We refer to the first three methods as AFT–Zellner, AFT–pMOMZ and AFT–LASSO. They all assume an AFT model and use either the block-Zellner prior π_L, the nonlocal pMOM–Zellner prior π_M (Section 2.3), or LASSO penalties as proposed by [26]. AFT–Zellner and AFT–pMOMZ assume the Normal AFT model in (1), whereas AFT–LASSO uses a semiparametric AFT model. The remaining two methods combine the Cox model with piMOM priors (Cox–piMOM, [33]) and LASSO (Cox–LASSO, [48]). For AFT–Zellner and AFT–pMOMZ we used the function `modelSelection` in the R package `mombf` with the default prior parameters, the Beta-Binomial prior $\pi(\gamma)$ in (6) and $B = 10,000$ iterations in Algorithm 1. For Cox–piMOM we the used function `cov_bvs` in the R package `BMSNLP` with default parameters and prior dispersion 0.25 as recommended by [33]. For AFT–LASSO and Cox–LASSO we used the functions `AEnet.aft` and `glmnet` in the R packages

`AdapEnetClass` and `glmnet`, and we set the penalization parameter via 10-fold cross-validation.

5.1 Censoring, Model Complexity and Misspecification with $p = 2$

We consider sample sizes $n \in \{100, 500\}$, as well as censored and uncensored data. We present results for AFT–pMOMZ, as those for AFT–Zellner and Cox–piMOM were largely analogous. These methods are compared to Cox–LASSO and AFT–LASSO in Section 5.2. We consider six simulation scenarios. Scenarios 1–2 have a data-generating AFT model, Scenarios 3–4 a generalized hazard model and Scenarios 5–6 a proportional hazards model. The first covariate has a linear effect in all scenarios, whereas the second covariate has a nonlinear effect. In Scenarios 1, 3 and 5 this effect is strongly nonlinear and nonmonotone, whereas in Scenarios 2, 4 and 6 it is monotone and can be roughly approximated by a linear trend; see Figure 1.

SCENARIO 1. *AFT structure with* $\log o_i = x_{i1} + 0.5\log(|x_{i2}|) + \epsilon_i$ *and* $c_i = 0.5$, *where* $x_i \sim N(0, A)$, $A_{11} = A_{22} = 1$, $A_{12} = 0.5$, $\epsilon_i \sim N(0, \sigma = 0.5)$.

SCENARIO 2. *AFT structure with* $\log o_i = x_{i1} + 0.5\log(1 + x_{i2}) + \epsilon_i$ *and* $c_i = 1$, *where* $x_i = (\tilde{x}_{i1}, |\tilde{x}_{i2}|)$, $\tilde{x}_i \sim N(0, A)$ *and* A, ϵ_i *as in Scenario* 1.

SCENARIO 3. *Generalized hazards structure with*

$$h_{GH}(t) = h_0\big(t\exp\{-x_{i1}/3 + 0.5\log(|x_{i2}|)\}\big) \times \exp\{-x_{i1}/3 + 0.75\log(|x_{i2}|)\},$$

$c_i = 0.5$, h_0 *being the Log-Normal*(0,0.5) *baseline hazard and* x_i *as in Scenario* 1.

SCENARIO 4. *Generalized hazards structure with*

$$h_{GH}(t) = h_0\big(t\exp\{-x_{i1}/3 + 0.5\log(1 + x_{i2})\}\big) \times \exp\{-x_{i1}/3 + 0.75\log(1 + x_{i2})\},$$

$c_i = 1$, *and* h_0 *and* x_i *as in Scenario* 3.

SCENARIO 5. *Proportional hazards with* $h(t) = h_0(t)\exp\{3x_{i1}/4 - 5\log(|x_{i2}|)/4\}$, $c_i = 0.55$, h_0 *being the Log-Normal*(0,0.5) *baseline hazard and* x_i *as in Scenario* 1.

SCENARIO 6. *Proportional hazards with* $h(t) = h_0(t)\exp\{3x_{i1}/4 - 5\log(|x_{i2}|)/4\}$, $c_i = 0.95$, *and* h_0 *and* x_i *as in Scenario* 5.

In all scenarios, we first consider that there is no censoring, and then a strong administrative censoring, giving censoring probabilities $P_{F_0}(u_i = 0) \approx 0.7$.

We first discuss Scenarios 1–2 and illustrate the advantage of using our nonlinear effect decomposition. We first only considered the selection of nonlinear effects, that is, $\gamma_j \in \{0, 2\}$. In such case, the power to detect the effects (Figure S3, top) was significantly lower than when decomposing them into linear and nonlinear parts (Figure S3, middle). These findings align with Proposition 3, in the sense that the improvement in model fit needs to overcome the penalty for using a nonlinear basis. By considering $\gamma_j \in \{0, 1, 2\}$, one can capture part of the effect with a single linear term. Figure S3 also shows that censoring tends to reduce the power for both covariates.

Second, we illustrate the effect of the nonlinear basis dimension r. We compared the earlier results, where r was part of the model selection, to those obtained under a single fixed $r = 5$, 10 or 15 (Figure S3, bottom). Interestingly, in Scenario 1 the best performance was observed for $r = 5$, despite the data-generating truth being strongly nonlinear (Figure 1). In Scenario 2, the results were highly robust to r, as one might expect from the true effect being near-linear. That is, the smaller $r = 5$ gave a good compromise between inference and computation, we thus used $r = 5$ from now on.

The results for Scenarios 3–4 are in Figure S4, and for Scenarios 5–6 in Figure S5. The effect of censoring, model complexity and misspecifiying covariate effects were largely analogous to Scenarios 1–2. To explore further the effects of misspecification, we repeated the simulations in Scenarios 1–2 but now setting F_0 to have asymmetric Laplace errors $\epsilon_i \sim \text{ALaplace}(0, s, a)$, where $a = -0.5$ is the asymmetry and s the scale in the parameterization of [39]. We set s such that the error variance was equal to the Normal simulations, that is, $s = \sigma^2/[2(1 + a^2)] = 0.1$. Figure S6 shows the results. These are similar to Figure S3 except for a slight drop in the power to include active covariates.

Finally, we explored the effect of omitting covariates by analyzing the data from Scenarios 1–2 but considering that only x_{i1} was actually observed, that is, removing x_{i2} from the analysis. Figure S7 shows the results. Relative to Figure S3, under Scenario 1 there was a reduction in the posterior evidence for including x_{i1}. Such reduction was not observed in Scenario 2, presumably due to x_{i1} being correlated with $\log(1 + x_{i2})$ and hence picking up part of its predictive power.

5.2 Censoring, Model Complexity and Misspecification with $p = 50$

We extended Scenarios 1–6 from Section 5.1 by adding 48 spurious covariates. We generated covariates $x_i \sim N(0, A)$ where A is a 50×50 matrix with unit diagonal and all off-diagonal $A_{ij} = 0.5$, and otherwise simulated data as in Section 5.1. Figure 2 shows the proportion of correct model selections by each model selection method in Scenarios 1–2, across 250 independent simulations. Figure S8 reports these results for Scenarios 3–4, and Figure S9 for Scenarios 5–6. Tables S1–S6 also display

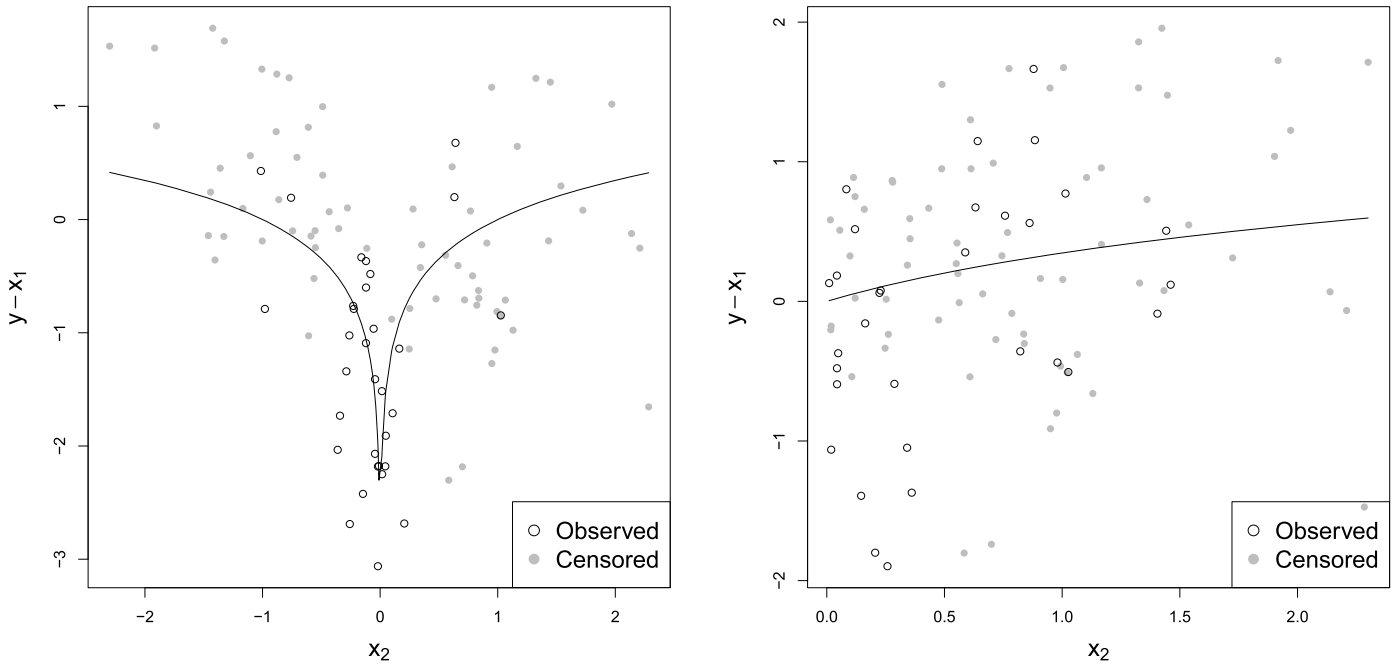

FIG. 1. *Simulation truth and a simulated dataset for Scenarios* 1 (*left*) *and* 2 (*right*).

the posterior probability assigned to the optimal model γ^* and the average number of truly active and truly inactive selected covariates. All Bayesian methods exhibited a good ability to select γ^* that improved with larger n and uncensored data (as predicted by Proposition 3), and they all provided significant improvements over Cox–LASSO and AFT–LASSO, particularly in reducing the number of false positives. As expected AFT–Zellner and AFT–pMOM tended to slightly outperform Cox–piMOM under truly AFT data (Scenarios 1—2), and conversely under truly proportional hazards data (Scenarios 5—6), though the differences were relatively minor. Interestingly, under the generalized hazards model (Scenarios 3–4) again AFT–Zellner and AFT–pMOMZ achieved higher correct selection rates, presumably due to these generalized hazard settings being closer to an AFT than to an proportional hazards model.

5.3 Effect of TGFB and Fibroblasts in Colon Cancer Metastasis

[4] studied the effect of 172 genes related to fibroblasts (f-TBRS signature), a cell type producing the structural framework in animals, and a growth factor (TGFB) associated with lower colon cancer survival (time until recurrence). The authors obtained 172 genes responsive to TGFB in mice fibroblasts. They then used independent gene expression data from human patients, with tumor stages 1-3, to show that an overall high mean expression of these 172 genes was strongly associated with metastasis. We analyzed their data to provide a more detailed description of the role of TGFB and f-TBRS on survival. We used the $n = 260$ patients with available survival times, and used tumor stage (two dummy indicators), TGFB and the 172 f-TBRS genes as covariates, for a total of $p = 175$. We first performed model selection via AFT–pMOMZ only for staging and TGFB. The top model had 0.976 posterior probability and included stage and a linear effect of TGFB, confirming that TGFB is associated with metastasis. The posterior marginal inclusion probability for a nonlinear effect of TGFB was only 0.009. As an additional check, the maximum likelihood estimator under the top model gave P-values < 0.001 for stage and the linear TGFB effect. The estimated time accelerations associated to TGFB are substantial (Figure S12, left).

Next, we extended the exercise to all 175 variables, only considering linear effects. The top model contained gene FLT1 and the second top model genes ESM1 and GAS1, with respective posterior model probabilities 0.088 and 0.081. These were also the genes with highest inclusion probabilities (0.208, 0.699 and 0.567 respectively). There is plausible biology connecting FLT1, ESM1 and GAS1 to metastasis. From genecards.org [50], FLT1 is a growth and permeability factor in cell proliferation and cancer invasion. ESM1 is related to endothelium disorders, growth factor receptor binding and gastric cancer networks, and GAS1 plays a role in growth and tumor suppression. Interestingly the marginal inclusion probability for TGFB was only 0.107, that is, after accounting for the top 3 genes TGFB did not show a significant effect on survival. For confirmation, we fitted via maximum likelihood the model with FLT1, ESM1, GAS1, stage and TGFB. The P-value for TGFB was 0.281 and its estimated effect was substantially reduced (Figure S12, right). Finally, we considered both linear and nonlinear effects ($p(1+r) = 1050$

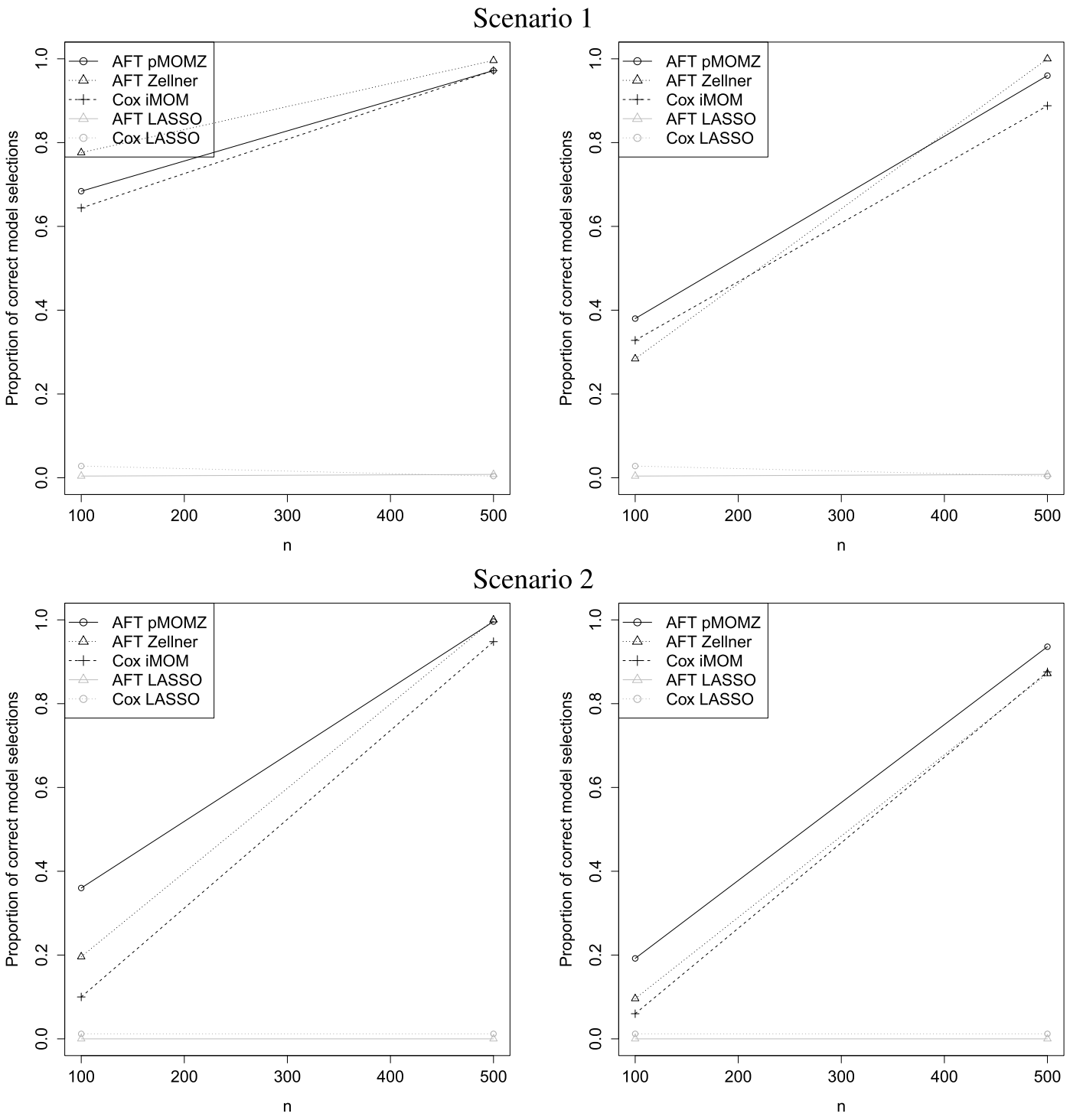

FIG. 2. *Scenarios* 1–2, $p = 50$. *Correct model selection proportion in uncensored* (*left*) *and censored* (*right*) *data.*

columns in (X, S)). All nonlinear effects had inclusion probabilities below 0.5 and the top 2 models contained FLT1, ESM1 and GAS1, as before. For comparison we run Cox–piMOM, AFT—LASSO and Cox–LASSO on the linear effects ($p = 175$). Stage and FLT1 were again selected by the top model under Cox–piMOM and by Cox–LASSO. Cox–LASSO selected 9 other genes, but only 4 had a significant P-value upon fitting a Cox model via maximum likelihood. Finally AFT–LASSO selected stage and six genes, two of which were also selected by Cox–LASSO. See Section S11.3 for a similar analysis of the estrogen receptor ESR1 effect on breast cancer.

Since this is a real-data application with an unknown ground truth, it is hard to assess which method performed best. As a first check, Table S7 reports the estimated predictive accuracy of each method via the leave-one-out cross-validated concordance index [15]. Cox–LASSO and AFT–pMOMZ achieved the highest concordance indexes,

with the former selecting more variables than the latter on average across the cross-validation (13.6 vs. 3.9 for $p = 175$ and 11.7 vs. 4.9 for $p(1 + r) = 1050$). We remark that predictive accuracy is not our primary goal, but if a method were to miss truly active covariates then one would expect accuracy to decrease, hence it serves as a rough proxy for statistical power. To complete the exercise, we next evaluate false positive probabilities.

5.4 False Positive Assessment Under Colon Cancer Data

We did a permutation exercise to assess false positive findings in the colon cancer data. We randomly permuted the recurrence times and left the covariates unpermuted. We obtained 100 independent permutations and recorded the model selected by each method. We first included only stage, a linear and nonlinear term for TGFB as covariates, for a total of $p(r + 1) = 8$ columns in (X, S). Next, we repeated the exercise considering linear effects for staging and the 173 genes, for a total of $p = 175$ columns.

The results are in Table 1 and Figure S10. AFT–pMOMZ achieved an excellent false positive control, it selected the null model in all permutations and assigned an average posterior probability $\pi(\gamma = 0 \mid y) = 0.846$ and 0.844 to the null model in the exercises with eight and 175 columns (respectively). That is, AFT–pMOMZ not only selected the null model but also assigned a high confidence to that selection. All competing methods selected the null model significantly less frequently. They also showed inflated false positive percentages for the analysis with 8 columns, though interestingly these percentages were lower in the analysis with 175 columns. Figure S10 reveals an interesting pattern for Cox–piMOM, in $> 97\%$ of the permutations only one covariate was included. That is, although the mean false positives percentage for Cox–piMOM was similar to AFT–LASSO and Cox–LASSO, the selected model was always very close to the null model, as expected from the strong sparsity-inducing properties of nonlocal priors.

TABLE 1
Percentage of false positives and correct model selections ($\widehat{\gamma} = 0$) *in permuted colon cancer data* (100 *permutations*) *when the design had eight columns* (*stage*, *linear and nonlinear effect of TGFB*) *and* 175 *columns* (*stage and linear effect of* 173 *genes*)

	Stage + TGFB ($p(r + 1) = 8$)		Stage + all genes ($p = 175$)	
	False positives	$\widehat{\gamma} = 0$	False positives	$\widehat{\gamma} = 0$
AFT–pMOMZ	0.0	100.0	0.0	100.0
Cox–piMOM	12.1	3.0	0.6	1.0
AFT–LASSO	35.9	31.0	2.2	45.0
Cox–LASSO	12.6	68.0	1.5	61.0

6. DISCUSSION

Our main contributions are describing a generic Bayesian model selection framework to incorporate nonlinear effects in a data-driven fashion to balance power and sparsity and, perhaps more importantly, helping understand the interplay between censoring, misspecification and model complexity. In survival models, we showed that one asymptotically discards covariates that do not help predict the outcome neither censoring times (conditionally on other covariates), whereas in probit regression one keeps those that help reduce the probit loss function, and similarly for other concave log-likelihoods. We showed that censoring and misspecification can reduce power significantly. Understanding this phenomenon can be useful in the design of experiments, where one may increase the follow-up length to gain power. Enriching the model class, by considering semi- and nonparametric terms, to alleviate model misspecification requires some care as these additional terms can incur computational and statistical power losses. Our recommendation is to use Bayesian model selection to decide their inclusion in a data-adaptive manner, as in the proposed linear plus deviation from linearity decomposition. Although not discussed here for simplicity, one can also easily incorporate interactions between covariates into the proposed theory and computational methods.

From a technical point of view we used standard asymptotic arguments which, for concave log-likelihoods, lead to simpler proofs and technical conditions. It should be possible to extend our results, with some care, to nonconcave and nonasymptotic settings (e.g., using the high-dimensional framework in [34]), interval and left censored data, as well as to cure rate, recurrence or excess hazards models. We focused on fixed p to provide simpler results and intuition, under less restrictive technical conditions. While, in theory, it can be potentially interesting to allow the nonlinear basis dimension r to grow with n, for actual methodology this often implies an impractical computational cost. This is critical in structural learning, where one wishes to consider many models. For this reason, in applied settings, it is common to use a finite basis.

Regarding high-dimensional settings, from recent results on misspecified penalized nonconcave likelihood [31] Bayesian model selection [37, 56], we speculate that our main findings should remain valid. We remark, however, that high-dimensional formulations often incorporate stronger sparsity via the prior distribution, hence the power drop caused by censoring and misspecification could be more problematic than in our fixed p case.

We focused on model selection within additive models, but our results extend directly when one wishes to consider interactions, by adding the corresponding basis

to our formulation. Our theory is valid for any given basis and also when performing selection on the basis itself, however, admittedly our examples focused on spline basis with fixed knots. We feel that a detailed study of basis selection would obscure the high-level intuition of our main results, but it represents an interesting aspect for future research.

ACKNOWLEDGMENTS

The authors thank Natalia Bochkina for pointing out an error in our original proof of Proposition 1 and suggesting a remedy that led to Assumption A3.

David Rossell is also affiliated to the Barcelona School of Economics, Data Science Center, Barcelona, Spain

FUNDING

David Rossell was supported by Spanish Government grants RyC-2015-18544, Plan Estatal PGC2018-101643-B-I00, Europa Excelencia EUR2020-112096 by the AEI/10.13039/501100011033 and European Union "NextGenerationEU"/PRTR, Ayudas Fundación BBVA a Investigación en Big Data 2017, and NIH grant R01 CA158113-01.

SUPPLEMENTARY MATERIAL

Supplementary Derivations, Proofs and Results (DOI: 10.1214/21-STS846SUPPA; .pdf). Likelihood and prior derivations, prior elicitation, computational algorithms, proofs of all propositions, supplementary figures and tables

Supplementary R Code and Data (DOI: 10.1214/21-STS846SUPPB; .zip). File supplementary_rcode.zip has the R code to reproduce the simulations, the F-TBRS analysis and the F-TBRS cancer data.

REFERENCES

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. *National Bureau of Standards Applied Mathematics Series*, *No*. 55. U.S. Government Printing Office, Washington, D.C. MR0167642

[2] BOCHKINA, N. A. and GREEN, P. J. (2014). The Bernstein-von Mises theorem and nonregular models. *Ann*. *Statist*. **42** 1850–1878. MR3262470 https://doi.org/10.1214/14-AOS1239

[3] BURRIDGE, J. (1981). A note of maximum likelihood estimation for regression models using grouped data. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **43** 41–45. MR0610375

[4] CALON, A., ESPINET, E., PALOMO-PONCE, S., TAURIELLO, D. V. F., IGLESIAS, M., CÉSPEDES, M. V., SEVILLANO, M., NADAL, C., JUNG, P. et al. (2012). Dependency of colorectal cancer on a TGF-beta-driven programme in stromal cells for metastasis initiation. *Cancer Cell* **22** 571–584.

[5] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *Ann*. *Statist*. **43** 1986–2018. MR3375874 https://doi.org/10.1214/15-AOS1334

[6] CHEN, Y. Q. and JEWELL, N. P. (2001). On a general class of semiparametric hazards regression models. *Biometrika* **88** 687–702. MR1859402 https://doi.org/10.1093/biomet/88.3.687

[7] CONSONNI, G., FOUSKAKIS, D., LISEO, B. and NTZOUFRAS, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Anal*. **13** 627–679. MR3807861 https://doi.org/10.1214/18-BA1103

[8] COX, D. R. (1972). Regression models and life-tables. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **34** 187–220. MR0341758

[9] DIRIENZO, A. G. and LAGAKOS, S. W. (2001). Effects of model misspecification on tests of no randomized treatment effect arising from Cox's proportional hazards model. *J*. *R*. *Stat*. *Soc*. *Ser*. *B*. *Stat*. *Methodol*. **63** 745–757. MR1872064 https://doi.org/10.1111/1467-9868.00310

[10] DUNSON, D. B. and HERRING, A. H. (2005). Bayesian model selection and averaging in additive and proportional hazards models. *Lifetime Data Anal*. **11** 213–232. MR2158783 https://doi.org/10.1007/s10985-004-0384-x

[11] FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics* **54** 1475–1485. MR1671590 https://doi.org/10.2307/2533672

[12] GASULL, A. and UTZET, F. (2014). Approximating Mills ratio. *J*. *Math*. *Anal*. *Appl*. **420** 1832–1853. MR3240110 https://doi.org/10.1016/j.jmaa.2014.05.034

[13] GRIFFIN, J. E., ŁATUSZYŃSKI, K. G. and STEEL, M. F. J. (2021). In search of lost mixing time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large *p*. *Biometrika* **108** 53–69. MR4226189 https://doi.org/10.1093/biomet/asaa055

[14] HAHN, P. R. and CARVALHO, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *J*. *Amer*. *Statist*. *Assoc*. **110** 435–448. MR3338514 https://doi.org/10.1080/01621459.2014.993077

[15] HARRELL JR., F. E., LEE, K. L. and MARK, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat*. *Med*. **15** 361–387.

[16] HATTORI, S. (2012). Testing the no-treatment effect based on a possibly misspecified accelerated failure time model. *Statist*. *Probab*. *Lett*. **82** 371–377. MR2875225 https://doi.org/10.1016/j.spl.2011.10.016

[17] HJORT, N. L. (1992). On inference in parametric survival data models. *Int*. *Stat*. *Rev*. **60** 355–387.

[18] HJORT, N. L. and POLLARD, D. (2011). Asymptotics for minimisers of convex processes. Available at arXiv:1107.3806.

[19] HOUGAARD, P. (1995). Frailty models for survival data. *Lifetime Data Anal*. **1** 255–273.

[20] HUANG, J., MA, S. and XIE, H. (2006). Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics* **62** 813–820. MR2247210 https://doi.org/10.1111/j.1541-0420.2006.00562.x

[21] HUTTON, J. L. and MONAGHAN, P. F. (2002). Choice of parametric accelerated life and proportional hazards models for survival data: Asymptotic results. *Lifetime Data Anal*. **8** 375–393. MR1942343 https://doi.org/10.1023/A:1020570922072

[22] IBRAHIM, J. G., CHEN, M.-H. and MACEACHERN, S. N. (1999). Bayesian variable selection for proportional hazards models. *Canad*. *J*. *Statist*. **27** 701–717. MR1767142 https://doi.org/10.2307/3316126

[23] JOHNSON, V. E. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 143–170. MR2830762 https://doi.org/10.1111/j.1467-9868.2009.00730.x

[24] JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *J. Amer. Statist. Assoc.* **107** 649–660. MR2980074 https://doi.org/10.1080/01621459.2012.682536

[25] KEIDING, N., ANDERSEN, P. K. and KLEIN, J. P. (1997). The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Stat. Med.* **16** 215–224.

[26] KHAN, M. H. R. and SHAW, J. E. H. (2019). Variable selection for accelerated lifetime models with synthesized estimation techniques. *Stat. Methods Med. Res.* **28** 937–952. MR3922901 https://doi.org/10.1177/0962280217739522

[27] LAIRD, N. and OLIVIER, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *J. Amer. Statist. Assoc.* **76** 231–240. MR0624329

[28] IBRAHIM, J. G. and CHEN, M. H. (2014). Bayesian model selection in survival analysis. In *Wiley StatsRef: Statistics Reference Online*. American Cancer Society.

[29] LEE, C.-I. C. (1992). On Laplace continued fraction for the normal integral. *Ann. Inst. Statist. Math.* **44** 107–120. MR1165575 https://doi.org/10.1007/BF00048673

[30] LIN, D. Y. and WEI, L. J. (1989). The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.* **84** 1074–1078. MR1134495

[31] LOH, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45** 866–896. MR3650403 https://doi.org/10.1214/16-AOS1471

[32] NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.* **42** 789–817. MR3210987 https://doi.org/10.1214/14-AOS1207

[33] NIKOOIENEJAD, A., WANG, W. and JOHNSON, V. E. (2020). Bayesian variable selection for survival data using inverse moment priors. *Ann. Appl. Stat.* **14** 809–828. MR4117831 https://doi.org/10.1214/20-AOAS1325

[34] PANOV, M. and SPOKOINY, V. (2015). Finite sample Bernstein–von Mises theorem for semiparametric problems. *Bayesian Anal.* **10** 665–710. MR3420819 https://doi.org/10.1214/14-BA926

[35] POLSON, N. G. and SUN, L. (2019). Bayesian l_0-regularized least squares. *Appl. Stoch. Models Bus. Ind.* **35** 717–731. MR3974246 https://doi.org/10.1002/asmb.2381

[36] ROSEN, J. B. (1971). Minimum error bounds for multidimensional spline approximation. *J. Comput. System Sci.* **5** 430–452. MR0283462 https://doi.org/10.1016/S0022-0000(71)80026-0

[37] ROSSELL, D. (2021). A framework for posterior consistency in model selection. *Bayesian Anal.* **in press**.

[38] ROSSELL, D., ABRIL, O. and BHATTACHARYA, A. (2021). Approximate Laplace approximations for scalable model selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **83** 853–879. MR4320004

[39] ROSSELL, D. and RUBIO, F. J. (2018). Tractable Bayesian variable selection: Beyond normality. *J. Amer. Statist. Assoc.* **113** 1742–1758. MR3902243 https://doi.org/10.1080/01621459.2017.1371025

[40] ROSSELL, D. and RUBIO, F. J. (2023). Supplement to "Additive Bayesian variable selection under censoring and misspecification." https://doi.org/10.1214/21-STS846SUPPA, https://doi.org/10.1214/21-STS846SUPPB

[41] ROSSELL, D. and TELESCA, D. (2017). Nonlocal priors for high-dimensional estimation. *J. Amer. Statist. Assoc.* **112** 254–265. MR3646569 https://doi.org/10.1080/01621459.2015.1130634

[42] ROSSELL, D., TELESCA, D. and JOHNSON, V. E. (2013). High-dimensional Bayesian classifiers using non-local priors. In *Statistical Models for Data Analysis XV* 305–314. Springer, Berlin.

[43] SANDERSON, C. and CURTIN, R. (2016). Armadillo: A template-based C ++ library for linear algebra. *J. Open Sour. Softw.* **1** 26.

[44] SCHEIPL, F., FAHRMEIR, L. and KNEIB, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *J. Amer. Statist. Assoc.* **107** 1518–1532. MR3036413 https://doi.org/10.1080/01621459.2012.737742

[45] SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38** 2587–2619. MR2722450 https://doi.org/10.1214/10-AOS792

[46] SHA, N., TADESSE, M. G. and VANNUCCI, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22** 2262–2268.

[47] SILVAPULLE, M. J. and BURRIDGE, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *J. Roy. Statist. Soc. Ser. B* **48** 100–106. MR0848055

[48] SIMON, N., FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39** 1–13. https://doi.org/10.18637/jss.v039.i05

[49] SOLOMON, P. J. (1984). Effect of misspecification of regression models in the analysis of survival data. *Biometrika* **71** 291–298. MR0767157 https://doi.org/10.1093/biomet/71.2.291

[50] STELZER, G., ROSEN, N., PLASCHKES, I., ZIMMERMAN, S., TWIK, M., FISHILEVICH, S., STEIN, T. I., NUDEL, R., LIEDER, I. et al. (2016). The GeneCards suite: From gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics* **54** 1–30.

[51] STRUTHERS, C. A. and KALBFLEISCH, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73** 363–369. MR0855896 https://doi.org/10.1093/biomet/73.2.363

[52] TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.

[53] TONG, X., ZHU, L., LENG, C., LEISENRING, W. and ROBISON, L. L. (2013). A general semiparametric hazards regression model: Efficient estimation and structure selection. *Stat. Med.* **32** 4980–4994. MR3127189 https://doi.org/10.1002/sim.5885

[54] TSIATIS, A. A. (1981). A large sample study of Cox's regression model. *Ann. Statist.* **9** 93–108. MR0600535

[55] WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, New York.

[56] YANG, Y. and PATI, D. (2017). Bayesian model selection consistency and oracle inequality with intractable marginal likelihood. Available at arXiv:1701.00311.

[57] YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.* **44** 2497–2532. MR3576552 https://doi.org/10.1214/15-AOS1417

[58] YING, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21** 76–99. MR1212167 https://doi.org/10.1214/aos/1176349016

[59] ZANELLA, G. and ROBERTS, G. (2019). Scalable importance tempering and Bayesian variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **81** 489–517. MR3961496

[60] ZHANG, Z., SINHA, S., MAITI, T. and SHIPP, E. (2018). Bayesian variable selection in the accelerated failure time model with an application to the surveillance, epidemiology, and end results breast cancer data. *Stat. Methods Med. Res.* **27** 971–990. MR3770130 https://doi.org/10.1177/0962280215626947