Probability in the Engineering and Informational Sciences, **34**, 2020, 364–380.
doi:10.1017/S0269964818000190

VARIABLE SELECTION FOR BAYESIAN SURVIVAL MODELS USING BREGMAN DIVERGENCE MEASURE

DAOYUAN SHI AND LYNN KUO
Department of Statistics, University of Connecticut, Storrs, Connecticut, USA
E-mail: lynn.kuo@uconn.edu

The variable selection has been an important topic in regression and Bayesian survival analysis. In the era of rapid development of genomics and precision medicine, the topic is becoming more important and challenging. In addition to the challenges of handling censored data in survival analysis, we are facing increasing demand of handling big data with too many predictors where most of them may not be relevant to the prediction of the survival outcome. With the desire of improving upon the accuracy of prediction, we explore the Bregman divergence criterion in selecting predictive models. We develop sparse Bayesian formulation for parametric regression and semiparametric regression models and demonstrate how variable selection is done using the predictive approach. Model selections for a simulated data set, and two real-data sets (one for a kidney transplant study, and the other for a breast cancer microarray study at the Memorial Sloan-Kettering Cancer Center) are carried out to illustrate our methods.

Keywords: reliability theory, stochastic modeling

1. INTRODUCTION

Both Bayesian parametric and semiparametric regression methods for survival data have been developed extensively. The parametric regression models include exponential, Weibull, gamma, log-normal distributions for the survival times, where the mean of the survival time is modeled by a function (e.g., exponential) of the linear predictors. The semiparametric models include the Cox [5] proportional hazards model where the hazard function is assumed to be proportional to the baseline hazard function with the ratio to be the exponential transformation of the linear predictors. In this paper, we concentrate on the variable selection problem with modern genomic data where the number of predictors (p) is much larger than the sample size (n).

In variable selection with genomic data, the challenges are that the traditional subset regression and stepwise regression do not work anymore, due to too many predictors and too few samples. We seek sparse statistical models where there is only a small number of non-zero regression coefficients. The models are easier to estimate and interpret than dense models and provide reproducible and useful information with good predictive accuracy. Tibshirani [29] proposed least absolute shrinkage and selection operator (LASSO) and Gui and Li [14] proposed least angle regression LASSO procedure to the partial likelihood based on the

© Cambridge University Press 2018 0269-9648/18 $25.00

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

Cox proportional hazards model and showed they have much better predictive performance than the L_2 penalized regression by a simulation study. LASSO uses the L_1 penalty for regulation, so in the "large p, small n" case, LASSO selects at most n variables before it saturates. If there is a group of highly correlated variables, then LASSO tends to select one variable from a group and ignore the others. To overcome these limitations, Zou and Hastie [32] proposed the elastic net which adds L_2 penalty for regulation. So the elastic net method applied to survival analysis results in finding $\boldsymbol{\beta}$ that minimizes the penalized log partial likelihood:

$$-\ell(\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2,$$

where $\ell(\boldsymbol{\beta})$ denotes the log partial likelihood function of $\boldsymbol{\beta}$ proposed by Cox [6], and λ_1 and λ_2 are tuning parameters with λ_1 controlling sparsity and λ_2 controlling grouping behavior. The elastic net method includes LASSO and ridge regression as special cases. For the linear regression with normal error terms, the LASSO estimators can be interpreted as Bayesian estimators by choosing independent Laplacian prior for the regression coefficients (Park and Casella [26]). Similarly, the elastic net estimate can be related to the posterior mode where a composite of Laplacian and Gaussian prior is used for the regression coefficients to mimic the L_1 and L_2 penalties in the loss function (Li and Lin [25]). On the Bayesian LASSO version for the Cox regression model, Lee, Chakraborty and Sun [21] proposed a shrinkage prior on the regression coefficients through a scale mixture representation of normal and gamma distributions and the cumulative baseline hazard function is estimated using a discrete gamma process. Lee, Chakraborty and Sun [22] further developed a Bayesian version for the elastic net, fused lasso, and group lasso penalty by modifying the prior on the regression coefficients considered by Kyung et al. [20] earlier. They further developed Markov chain Monte Carlo (MCMC) algorithms for posterior inference.

The above variable selection problem is considered from the Bayesian inference point of view. In precision medicine, predictive accuracy is also important in addition to the ability to select prognostic markers for improved disease screening and prevention. So we pursue this further by concentrating on the predictive power of a selected model. The use of predictive distribution has played a major role in model selection in Bayesian inference. Stone [28] and Geisser and Eddy [8] proposed the pseudo-Bayes factor (log pseudo-marginal likelihood, LPML) based on the conditional predictive density approach using a cross-validation idea. In order to generalize the predictive model selection, we employ the Bregman divergence (BD) and efficient algorithms for minimizing it. BD [4], which defines a family of generalized entropy measures by various convex functions, is used to measure prediction errors. This family includes squared error loss, logistic loss, Itakura-Saito distance, Mahalanobis distance, squared Euclidean distance, linex loss, and KL-divergence as given in Table 1 of Benerjee et al. [1]. New divergence measures can be generated by introducing new convex functions in the definition. The advantages of the prediction approaches have been studied by Guerrier and Victoria-Feser [13]. They find this prediction divergence criterion tends to select far smaller models with generally better or comparable out-of-sample predictive performance. It is well suited in "sparse" settings with correlated covariates, commonly encountered these days in applications.

The major contribution of this paper is to illustrate how to perform variable selection in sparse setting for survival analysis using BD measure. By simulation and real data analysis, we show BD is desirable in model selection as discussed in the last paragraph. We also compare the results to that using LPML and deviance information criterion (DIC) [27]. It suggests extra information can be obtained from using the BD measure.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

A review of the prediction for right censored survival models is given in Section 2. Predictive model selection with BD criterion (BDC) is discussed in Section 3. In Section 4, a simple simulated data set is used to compare the model selection results using BDC to that of other criteria (AIC, DIC, LPML, and WAIC). In Section 5, we apply our methods to two real data sets: one is about the survival times for 863 kidney transplant patients with three predictors: gender, race, and age; the other is about survival times for 153 ovarian cancer patients each with 2,647 gene expressions as covariates. Finally, we conclude with final remarks on future work.

2. BAYESIAN SURVIVAL MODELS FOR RIGHT-CENSORED DATA

2.1. Fully Parametric Models

2.1.1. The exponential model. The survival function and density of the exponential model are given by $S(t|\lambda) = P[T > t|\lambda] = \exp\{-\lambda t\}$ and $f(t|\lambda) = \lambda \exp\{-\lambda t\}$. For subject i, assume y_i is the observed survival time subject to right-censoring, and ν_i is its indicator for censoring, which is 1 for real observed survival time and 0 for censored time. Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)'$, then for the observed survival data denoted by $D = (n, \boldsymbol{y}, \boldsymbol{v})$ with n subjects, we can write the likelihood function of λ as

$$L(\lambda|D) = \prod_{i=1}^{n} f(y_i|\lambda)^{v_i} S(y_i|\lambda)^{(1-v_i)} = \lambda^d \exp\left\{-\lambda \sum_{i=1}^{n} y_i\right\},$$

where $d = \sum_{i=1}^{n} v_i$. If we consider the following conjugate prior for λ with hyper-parameters (α_0, λ_0), that is, $\lambda \sim$ gamma $\mathcal{G}(\alpha_0, \lambda_0)$:

$$\pi(\lambda|\alpha_0, \lambda_0) \propto \lambda^{\alpha_0 - 1} \exp\{-\lambda_0 \lambda\},$$

then the resulting posterior distribution of λ is given by

$$\pi(\lambda|D) \propto \lambda^{\alpha_0 + d - 1} \exp\left\{-\lambda(\lambda_0 + \sum_{i=1}^{n} y_i)\right\},$$

which is a $\mathcal{G}(\alpha_0 + d, \lambda_0 + \sum_{i=1}^{n} y_i)$ distribution. And the posterior predictive distribution of a future failure time y_{new} is given by

$$\begin{aligned}
\pi(y_{\text{new}}|D) &= \int_0^\infty f(y_{\text{new}}|\lambda)\pi(\lambda|D)d\lambda \\
&\propto \int_0^\infty \lambda^{\alpha_0 + d + 1 - 1} \exp\left\{-\lambda(y_{\text{new}} + \lambda_0 + \sum_{i=1}^{n} y_i)\right\} d\lambda \\
&\propto (\lambda_0 + \sum_{i=1}^{n} y_i + y_{\text{new}})^{-(\alpha_0 + d + 1)}.
\end{aligned}$$

The normalized posterior predictive distribution is given by

$$\pi(y_{\text{new}}|D) = (\alpha_0 + d)\left(\alpha_0 + \sum_{i=1}^{n} y_i\right)^{(\alpha_0 + d)} \Big/ \left(\lambda_0 + \sum_{i=1}^{n} y_i + y_{\text{new}}\right)^{(\alpha_0 + d + 1)}.$$

And the predictive distribution is known as an inverse beta distribution.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

To build a regression model, we introduce covariates through λ, and write $\lambda_i = \varphi(\boldsymbol{z}_i'\boldsymbol{\beta})$, where $\boldsymbol{z}_i$ is a $p \times 1$ vector of covariates including an intercept for the ith subject, $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\varphi(.)$ is a known function. A common form of φ is to take $\varphi(\boldsymbol{z}_i'\boldsymbol{\beta}) = \exp\{\boldsymbol{z}_i'\boldsymbol{\beta}\}$, we are led to the likelihood function

$$L(\boldsymbol{\beta}|D) = \exp\left\{\sum_{i=1}^{n} v_i \boldsymbol{z}_i'\boldsymbol{\beta}\right\} \exp\left\{-\sum_{i=1}^{n} y_i \exp(\boldsymbol{z}_i'\boldsymbol{\beta})\right\},$$

where $D = (n, \boldsymbol{y}, \boldsymbol{v}, Z)$ and Z is a $n \times p$ matrix with $\boldsymbol{z}_i'$ in the ith row. Common prior distributions for $\boldsymbol{\beta}$ include an improper uniform prior or a normal prior. Suppose, we specify a p-dimensional normal prior for $\boldsymbol{\beta}$, denoted by $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where $\boldsymbol{\mu}_0$ denotes the prior mean and $\boldsymbol{\Sigma}_0$ denotes the prior covariance matrix. Then the posterior distribution of $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|D) \propto L(\boldsymbol{\beta}|D)\pi(\boldsymbol{\beta}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),$$

where $\pi(\boldsymbol{\beta}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ is the multivariate normal density with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$.

2.1.2. The Weibull model. The Weibull density

$$f(t|\alpha, \mu) = \alpha t^{(\alpha-1)} \exp\{\mu - \exp(\mu)t^{\alpha}\}$$

with survival function $S(t|\alpha, \mu) = \exp\{-\exp(\mu)t^{\alpha}\}$ is perhaps the most widely used parametric survival model. We can write the likelihood function of (α, μ) based on right-censored data D as

$$L(\alpha, \mu|D) = \alpha^d \exp\left\{d\mu + \sum_{i=1}^{n}(v_i(\alpha-1)\log(y_i) - \exp(\mu)y_i^{\alpha})\right\}.$$

When α is assumed known, we can select a gamma prior as the conjugate prior for $\exp\{\mu\}$. No joint conjugate prior is available when (α, μ) are both assumed unknown. In this case, a typical joint prior specification is to take α and μ to be independent, with a gamma $\mathcal{G}(\alpha_0, \kappa_0)$ prior for α and a normal $N(\mu_0, \sigma_0^2)$ prior for μ. The joint posterior distribution of (α, λ) is given by

$$\begin{aligned}\pi(\alpha, \mu|D) &\propto L(\alpha, \mu|D)\pi(\alpha|\alpha_0, \kappa_0)\pi(\mu|\mu_0, \sigma_0^2)\\ &\propto \alpha^{\alpha_0+d-1}\\ &\quad \times \exp\left\{d\mu + \sum_{i=1}^{n}(v_i(\alpha-1)\log(y_i) - \exp(\mu)y_i^{\alpha}) - \kappa_0\alpha - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}.\end{aligned}$$

The joint posterior distribution of (α, μ) does not have a closed form, but it can be shown that the conditional posterior distributions $[\alpha|\mu, D]$ and $[\mu|\alpha, D]$ are log-concave, and Gibbs sampling is straightforward for this model.

To build the Weibull regression model, we introduce covariates through μ, and write $\mu_i = \boldsymbol{z}_i'\boldsymbol{\beta}$. Common prior distributions for $\boldsymbol{\beta}$ include the uniform improper prior and a normal prior. Assuming a $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ prior for $\boldsymbol{\beta}$ and a gamma prior for α, we are led to

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

the joint posterior

$$\pi(\boldsymbol{\beta}, \alpha|D) \propto \alpha^{\alpha_0+d-1} \exp\left\{ \sum_{i=1}^{n} (v_i \boldsymbol{z}_i'\boldsymbol{\beta} + v_i(\alpha-1)\log(y_i) - y_i^{\alpha}\exp(\boldsymbol{z}_i'\boldsymbol{\beta})) - \kappa_0\alpha - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_0)\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_0) \right\}.$$

2.2. Semiparametric Models

The Cox proportional hazards model is defined for a hazard function $h(t)$ for a subject with failure time t and a $1 \times p$ vector of possibly time-dependent covariates $\boldsymbol{x}(t)$:

$$h(t|\boldsymbol{x}(t)) = h_0(t)\exp\{\boldsymbol{x}(t)\boldsymbol{\beta}\},$$

where h_0 is the baseline hazard function and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients. Unless mentioned otherwise, we will present our methods for the special case of time-constant (fixed) covariates $\boldsymbol{x}(t) = \boldsymbol{x}$. Cox [5] proposed a partial likelihood which does not depend on h_0 for inference for the regression coefficient $\boldsymbol{\beta}$ with a frequentist approach. We consider Bayesian inference for this model, where the baseline h_0 is (i) a constant, (ii) a power function of t, or (iii) a piecewise constant function. So we will first discuss the likelihood and prior for Bayesian inference.

One of the most convenient and popular discrete time models for semiparametric survival analysis is the piecewise constant hazard model. To construct this model, we first construct a finite partition of the time axis, $0 < s_1 < s_2 < \cdots < s_J$, with $s_J > y_i$ for all $i = 1, 2, \ldots, n$. Thus, we have the J intervals $(0, s_1], (s_1, s_2], \ldots, (s_{J-1}, s_J]$. In the jth interval, we assume a constant baseline hazard $h_0(t) = \lambda_j$ for $t \in I_j = (s_{j-1}, s_j]$. Letting $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)$, we can write the likelihood function as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|D) = \prod_{i=1}^{n}\prod_{j=1}^{J} (\lambda_j \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\})^{\delta_{ij}v_i} \times \exp\left\{ -\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right\}, \quad \textbf{(2.1)}$$

where $\delta_{ij} = 1$ if the ith subject failed or was censored in the jth interval, and 0 otherwise. The indicator δ_{ij} is needed to properly define the likelihood function over the J intervals. This semiparametric model sometimes referred to as a piecewise exponential model, is quite useful, and general, and can accommodate various shapes of the baseline hazard function over the intervals. Moreover, we note that if $J = 1$, the model reduces to a parametric exponential model with failure rate parameter λ.

2.2.1. Gamma process prior. The gamma process is perhaps the most commonly used non-parametric prior process for the Cox model. The seminal paper by Kalbfleisch [18] describes the gamma process prior for the baseline cumulative hazard function $H_0(\cdot)$. The gamma process can be described as follows. Let $\mathcal{G}(\alpha, \lambda)$ denote the gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\lambda > 0$. Let $\alpha(t)$, $t \geq 0$, be an increasing left continuous function such that $\alpha = 0$, and let $Z(t)$, $t \geq 0$, be a stochastic process with the properties: (i) $Z(0) = 0$; (ii) $Z(t)$ has independent increments in disjoint intervals; and (iii)

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

for $t > s$, $Z(t) - Z(s) \sim \mathcal{G}(c(\alpha(t) - \alpha(s)), c)$, a gamma distribution with shape parameter $c(\alpha(t) - \alpha(s))$, and scale parameter c. Then the process $\{Z(t) : t \geq 0\}$ is called a gamma process and is denoted by $Z(t) \sim \mathcal{GP}(c\alpha(t), c)$, where $\alpha(t)$ is the mean of $Z(t)$, and c is a precision or confidence parameter about the prior mean $\alpha(t)$. So, we consider:

$$H_0 \sim \mathcal{GP}(c_0 H^*, c_0),$$

as a prior for the cumulative baseline hazard function $H_0(t)$. The prior mean (prior guess) $H^*(t)$ of the unknown function $H_0(t)$ is an increasing function, and H^* is often assumed to be a known parametric function with hyper-parameter vector γ_0. For example, $H^*(t) = \gamma_0 t$ corresponds to the constant γ_0 as a prior guess for baseline hazard. Similarly, $H^*(t) = \eta_0 t^{\kappa_0}$ corresponds to a Weibull as a prior guess, where $\gamma_0 = (\eta_0, \kappa_0)'$ is a vector of hyper-parameters. Let h_j denote the increment of the baseline hazard in the jth interval

$$h_j = H_0(s_j) - H_0(s_{j-1}), \quad j = 1, 2, \ldots, J.$$

The gamma process implies that the h_j follows independent gamma distribution, that is

$$h_j \sim \mathcal{G}(\alpha_{0j} - \alpha_{0,j-1}, c_0),$$

where $\alpha_{0j} = c_0 H^*(s_j)$.

2.2.2. Posterior distribution. The joint survival function of $T = (T_1, \ldots, T_n)$ for the Cox model is given by

$$P(\boldsymbol{T} > \boldsymbol{t} | \boldsymbol{\beta}, \boldsymbol{X}, H_0) = \exp\left\{ -\sum_{i=1}^{n} \exp(\boldsymbol{x}_i' \boldsymbol{\beta}) H_0(t_i) \right\},$$

where $H_0(t)$ is the cumulative baseline hazard function. The conditional probability of the ith subject failing in the interval I_j is given by

$$\begin{aligned} P(T_i \in I_j | \boldsymbol{h}) &= \exp\{-\exp(\boldsymbol{x}_i' \boldsymbol{\beta}) H(t_{j-1})\} - \exp\{\exp(\boldsymbol{x}_i' \boldsymbol{\beta}) H(t_i)\} \\ &= \exp\left\{ -\exp(\boldsymbol{x}_i' \boldsymbol{\beta}) \sum_{m=1}^{j-1} h_m \right\} [1 - \exp\{-h_j \exp(\boldsymbol{x}_i' \boldsymbol{\beta})\}]. \end{aligned}$$

Therefore, combining the likelihood function in Eq. (2.1) and the prior for $\boldsymbol{h}$ and $\boldsymbol{\beta}$, we have the posterior (see Ibrahim et al. [16]):

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{h} | D) \propto \prod_{j=1}^{J} \left(\exp\left\{ -h_j \sum_{k \in R_j - D_j} \exp(\boldsymbol{x}_k' \boldsymbol{\beta}) \right\} \prod_{\ell \in D_j} [1 - \exp\{-h_j \exp(\boldsymbol{x}_\ell' \boldsymbol{\beta})\}] \right) \\ \times \left(\prod_{j=1}^{J} h_j^{\alpha_{0j} - \alpha_{0,j-1} - 1} \exp(-c_0 h_j) \right) \pi(\boldsymbol{\beta}), \end{aligned} \tag{2.2}$$

where D_j is the set of subjects failing in the interval $I_j = (s_{j-1}, s_j]$ and the risk set R_j is the set of subjects "at risk" in the interval I_j.

If the data set has a "large p, small n" structure, then we will first consider the Bayesian LASSO formulation.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

Lee et al. [21] proposed the following hierarchical prior for $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots \tau_p^2 \sim N_p(0_p, \sigma^2 D_\tau), D_\tau = \text{diag}(\tau_1^2, \ldots, \tau_p^2),$$
$$\tau_1^2, \ldots, \tau_p^2|\lambda \sim \prod_{j=1}^{p} \frac{\lambda^2}{2} \exp\{-\lambda^2 \tau_j^2/2\} d\tau_j^2,$$
$$\pi(\sigma^2) = \frac{1}{\sigma^2} d\sigma^2.$$

The λ is a tuning parameter, which controls the sparsity of the model. Large values of λ induce higher sparsity. It is usually chosen from a cross-validation method in the frequentist's domain. From the Bayesian perspective, it can be chosen from an empirical Bayes approach. However, the computation can be intensive for large p. So we use hierarchical Bayes method assuming a diffuse hyper prior for λ and choose a diffuse gamma prior for λ^2:

$$\pi(\lambda^2) = \frac{\eta^\varphi}{\Gamma(\varphi)} (\lambda^2)^{\varphi - 1} \exp\{-\eta \lambda^2\},$$

where φ and η are the shape and the rate parameters of the gamma distribution.

For the Bayesian elastic net, Lee et al. [22] proposed the following hierarchical prior for $\boldsymbol{\beta}$.

$$\boldsymbol{\beta}|\sigma^2, D_\tau^* \sim N_p(0_p, \sigma^2 D_\tau^*), D_\tau^* = \text{diag}((\tau_1^{-2} + \lambda_2)^{-1}, \ldots, (\tau_p^{-2} + \lambda_2)^{-1}),$$
$$\tau_1^2, \ldots, \tau_p^2|\lambda_1^2, \lambda_2 \sim \prod_{j=1}^{p} \frac{1}{\sqrt{1 + \lambda_2 \tau_j^2}} \exp\{-\lambda_1^2 \tau_j^2\}, \pi(\sigma^2) = \frac{1}{\sigma^2} d\sigma^2.$$

In addition to Bayesian LASSO and elastic net, Lee et al. [22] has also proposed priors for Bayesian group lasso, and fused lasso respectively, and developed MCMC algorithms for all of them.

3. PREDICTIVE MODEL SELECTION WITH BDC

3.1. BD Measure

To measure predictive accuracy, mean squared error, weighted (by inverse variance) mean squared error, log predictive density, Akaike information criterion (AIC), DIC, LPML, Watanabe-AIC (WAIC) have been studied. We would like to add BD information criterion to measure the prediction accuracy.

Definition 1 (Bregman Divergence): *Let $\psi : \Omega \to R$ be a strictly convex and differentiable function on a non-empty convex set $\Omega \subseteq R^n$. Then, for $\boldsymbol{x}, \boldsymbol{y} \in R^n$, the BD_ψ is defined as*

$$BD_\psi(\boldsymbol{x}, \boldsymbol{y}) = \psi(\boldsymbol{x}) - \psi(\boldsymbol{y}) - \langle \boldsymbol{x} - \boldsymbol{y}, \nabla\psi(\boldsymbol{y}) \rangle,$$

where $\nabla\psi$ represents the gradient vector of ψ, and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is the inner product between the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

Figure 1 illustrates the BD measure for a convex function ψ between two points x and y in R^1.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

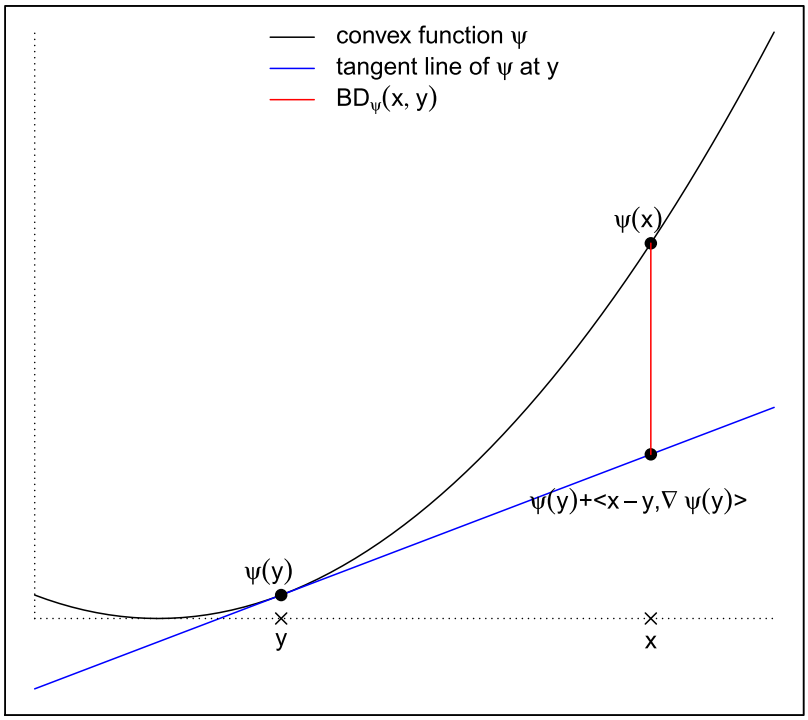


FIGURE **1.** $D_\psi(x, y) = \text{BD}$ for the function ψ between x and y . It is the distance marked in red between the tangent line and the convex function evaluated at $X = x$. In this example, the convex function ψ is chosen to be $\psi(z) = (z-1)^2$. The two points x and y are chosen to be 5 and 2 on the X-axis.

Table 1 of Banerjee et al. [1] exhibits various measures of BD by varying the choice of the ψ function. They include some well-known distance measures in the literature. In particular, Goh and Dey [12] exhibits the following three examples:

Example 1: Suppose $\psi(\boldsymbol{x}) = \sum_i x_i \log x_i$, then

$$BD_\psi(\boldsymbol{x}, \boldsymbol{y}) = \sum_i x_i \log \frac{x_i}{y_i} - x_i + y_i.$$

It is the generalized Kullback—Leibler divergence between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Example 2: Suppose $\psi(\boldsymbol{x}) = \sum_i -\log x_i$, then

$$BD_\psi(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \frac{x_i}{y_i} - \log \frac{x_i}{y_i} - 1.$$

It is the Itakura—Saito distance [17] between $\boldsymbol{x}$ and $\boldsymbol{y}$.

Example 3: Suppose $\psi(\boldsymbol{x}) = \boldsymbol{x}^T A \boldsymbol{x}$, where A is positive definite,

$$\begin{aligned} BD_\psi(\boldsymbol{x}, \boldsymbol{y}) &= \boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{y}^T A \boldsymbol{y} - \langle \boldsymbol{x} - \boldsymbol{y}, (A + A^T)\boldsymbol{y} \rangle \\ &= (\boldsymbol{x} - \boldsymbol{y})^T A (\boldsymbol{x} - \boldsymbol{y}). \end{aligned}$$

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

If A is assumed to be the inverse of the covariance matrix, then it is the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{y}$. However, if we assume $A = I$, then the BD reduces to the square Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$.

3.2. Predictive Distribution

BD can be used as a very general class of Bayesian model selection criteria. Let $\boldsymbol{y}_{s_i}$ denote a subset of $\boldsymbol{y}$, that is, $\boldsymbol{y}_{s_i} \subseteq \boldsymbol{y}$. Let $\boldsymbol{y}_{-s_i} = \boldsymbol{y} \backslash \boldsymbol{y}_{s_i}$. By choosing different subsets, we can relate our approach to a few well-known Bayesian model selection approaches as discussed in Gelfand and Dey [9], and Goh [11].

(a) If we choose all observations as the subset $\boldsymbol{y}_{s_i}$, that is, $\boldsymbol{y}_{s_i} = \boldsymbol{y}$ and naturally $\boldsymbol{y}_{-s_i} = \emptyset$, then we evaluate the standard marginal (or predictive) density in the evaluation of BD. Therefore, this approach reduces to the Bayes factor (BF) approach for Bayesian model selection.

(b) Let the subset $\boldsymbol{y}_{s_i}$ be a single observation, that is, $\boldsymbol{y}_{s_i} = y_i$ and $\boldsymbol{y}_{-s_i} = \boldsymbol{y} \backslash y_i$. In this case, we need to evaluate $p(y_i|\boldsymbol{y}_{-s_i}, M)$ for a model M, this is exactly the conditional predictive ordinate (CPO). Hence this approach reduces to the pseudo-BF (PSBF) approach, proposed by Geisser and Eddy [8].

(c) Suppose that $\{\boldsymbol{y}_{-s_i}\}$ is a minimal subset, that is $\{\boldsymbol{y}_{-s_i}\}$ includes the minimal number of observations such that the marginal or predictive density $m(\boldsymbol{y}_{-s_i})$ is proper under any model M. Thus it yields a generalization of the intrinsic BF (IBF) approach (Berger and Pericchi [2]).

(d) Assume that observations $y_1, y_2, \ldots, y_n$ are sequentially observed. Let $\boldsymbol{y}_{s_i}$ be ith observation and $\boldsymbol{y}_{-s_i}$ be all the observations prior to the ith observation such that $\boldsymbol{y}_{s_i} = y_i$ and $\boldsymbol{y}_{-s_i} = (y_1, y_2, \ldots, y_{i-1})$. Then this approach reduces to the prequential approach (Dawid 1984 [7]) to model selection.

3.3. Best Predictive Model

The predictive model selection problem can be considered as a decision-making problem. From this viewpoint, the best choice among proposed models is to select the model that provides the closest predictive distribution to the true distribution. Let $M \in \mathcal{M} = \{M^1, M^2, \ldots, M^K\}$ be a finite set of proposed models, where K is assumed to be known, and we assume one of the candidate models is the true model. This is referred as the $\mathcal{M}$ closed frame work (Barnardo and Smith [3]). Let $\boldsymbol{y}_{s_1}, \boldsymbol{y}_{s_2}, \ldots, \boldsymbol{y}_{s_m}$ be $m(\leq n)$ subsets of $\boldsymbol{y}$ such that $\boldsymbol{y} = (\boldsymbol{y}_{s_1}, \boldsymbol{y}_{s_2}, \ldots, \boldsymbol{y}_{s_m})$. Define $\boldsymbol{y}_{-s_i}$ to be a complement of $\boldsymbol{y}_{s_i}$, that is, $\boldsymbol{y}_{-s_i} = \boldsymbol{y} \backslash \boldsymbol{y}_{s_i}$ for $i = 1, 2, \ldots, m$. If we choose $\boldsymbol{y}_{s_i}$ to be y_i, then this yields the leave-one-out approach which is commonly used for small or moderate-sized data. For large data, we can use 10% of the non-overlapping data constructed sequentially, this leads to the 10-fold cross-validation.

Let M be the unknown true model (the unknown state of nature) and assume $M \in \mathcal{M}$. For a model $a \in \mathcal{M}$, let $L(M, a) = BD(p(\boldsymbol{y}_{\text{new}}|\boldsymbol{y}_{\text{old}}, M), p(\boldsymbol{y}_{\text{new}}|\boldsymbol{y}_{\text{old}}, a))$. Then the posterior expected loss is given as

$$\rho(\pi(M|\boldsymbol{y}_{\text{old}}), a) = \int_{M \in \mathcal{M}} L(M, a) d^{\pi(M|\boldsymbol{y}_{\text{old}})}(M), \tag{3.1}$$

where

$$\pi(M|\boldsymbol{y}_{\text{old}}) = \frac{m(\boldsymbol{y}_{\text{old}}|M)\pi(M)}{\int_{M \in \mathcal{M}} m(\boldsymbol{y}_{\text{old}}|M) d^{\pi(M)}(M)},$$

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

the prior $\pi(M)$ reflects our prior belief on the probability of the true model. We choose equally likely prior for the true model:

$$\pi(M) = \begin{cases} \frac{1}{K}, & M \in \mathcal{M} \\ 0, & M \notin \mathcal{M}. \end{cases}$$

So under this prior, the posterior expected loss in (3.1) reduces to

$$\rho(\pi(M|\boldsymbol{y}_{\text{old}}), a) = \frac{\sum_{\ell=1}^{K} BD_{\psi}(p(\boldsymbol{y}_{\text{new}}|\boldsymbol{y}_{\text{old}}, M^{\ell}), p(\boldsymbol{y}_{\text{new}}|\boldsymbol{y}_{\text{old}}, a)) m(\boldsymbol{y}_{\text{old}}|M^{\ell})}{\sum_{k=1}^{K} m(\boldsymbol{y}_{\text{old}}|M^{k})}.$$

Similarly, for each observed $\boldsymbol{y}_{s_i}$, we can compute the posterior expected BD loss for a model a:

$$\rho_{\psi}^{*}(\pi(M|\boldsymbol{y}_{-s_i}), a) = \frac{\sum_{\ell=1}^{K} BD_{\psi}(p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i}, M^{\ell}), p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i}, a)) m(\boldsymbol{y}_{-s_i}|M^{\ell})}{\sum_{k=1}^{K} m(\boldsymbol{y}_{-s_i}|M^{k})}.$$

Then, we define our best model by

$$\hat{M} = \underset{a \in \mathcal{M}}{\operatorname{argmin}} \sum_{i=1}^{m} \rho_{\psi}^{*}(\pi(M|\boldsymbol{y}_{-s_i}), a). \tag{3.2}$$

Next, we use $p_i^{(a)}$ to denote $p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i}, a)$ for model a, and $\boldsymbol{p}^{(a)} = (p_1^{(a)}, p_2^{(a)}, \ldots, p_m^{(a)})$. We define $\bar{\boldsymbol{p}} = (\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_m)$, where

$$\bar{p}_i = \sum_{\ell=1}^{K} \frac{m(\boldsymbol{y}_{-s_i}|M^{\ell})}{\sum_{k=1}^{K} m(\boldsymbol{y}_{-s_i}|M^{k})} p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i}, M^{\ell}) = \left\{ \sum_{\ell=1}^{K} m(\boldsymbol{y}|M^{\ell}) \right\} / \left\{ \sum_{t=1}^{K} m(\boldsymbol{y}_{-s_i}|M^{t}) \right\}.$$

Similarly, we use $\boldsymbol{p}^{(M^k)}$ to denote $(p_1^{(M^k)}, p_2^{(M^k)}, \ldots, p_m^{(M^k)})$, with $p_i^{(M^k)} = p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i}, M^k) = m(\boldsymbol{y}|M^k)/m(\boldsymbol{y}_{-s_i}|M^k)$.

THEOREM 1: *Bregman Divergnece Criterion*

$$\underset{a \in \mathcal{M}}{\operatorname{argmin}} \sum_{i=1}^{m} \rho_{\psi}^{*}(\pi(M|\boldsymbol{y}_{-s_i}), a) = \underset{a \in \mathcal{M}}{\operatorname{argmin}}\, \rho_{\psi}^{*}(\bar{\boldsymbol{p}}, \boldsymbol{p}^{(a)}) = \underset{k \in 1, \ldots, K}{\operatorname{argmin}}\, \rho_{\psi}^{*}(\bar{\boldsymbol{p}}, \boldsymbol{p}^{(M_k)}).$$

The proof is given in Theorem 1 of Goh and Dey [12], so it is omitted here.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

Note the BDC measures the distance between the predictive density vector of the model M^k and the average vector of predictive densities in the whole model space in the predictive space. So the model with the smallest BDC is preferred.

3.4. Calculation of CPO

According to the definition of BDC, it is also easily calculated as far as the CPO $p_i = p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i})$ is obtained. In some cases, fortunately, we can obtain marginal densities, $m(\boldsymbol{y}_{s_i}, \boldsymbol{y}_{-s_i})$ and $m(\boldsymbol{y}_{-s_i})$, in a closed form. In this situation, the p_i is given as

$$p_i = p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i}) = \frac{m(\boldsymbol{y}_{s_i}, \boldsymbol{y}_{-s_i})}{m(\boldsymbol{y}_{-s_i})} = \frac{m(\boldsymbol{y})}{m(\boldsymbol{y}_{-s_i})}.$$

However, in general, calculation of the CPO p_i is not an easy task in a Bayesian framework, even if it can be numerically calculated, it will demand a huge computational burden to obtain all densities for all subsets $\boldsymbol{y}_{s_i}$. Note for a fixed model M, we have the following identity (omitting M in the expression):

$$\begin{aligned} p_i = p(\boldsymbol{y}_{s_i}|\boldsymbol{y}_{-s_i}) &= \int f(\boldsymbol{y}_i|\boldsymbol{\beta})\pi(\boldsymbol{\beta}|\boldsymbol{y}_{-s_i})d\boldsymbol{\beta} \\ &= \left(\int \frac{1}{f(\boldsymbol{y}_{s_i}|\boldsymbol{\beta})}\pi(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta}\right)^{-1}. \end{aligned} \tag{3.3}$$

A proof for the last equality in Eq. (3.3) can be found in Lewis et al. [24]. So we propose a Monte Carlo (MC) estimator of the CPO based on only the MCMC sample obtained from the full posterior distribution and all possible CPOs are simultaneously computed from this MCMC sample. Because we already have the MCMC sample from the posterior distribution based on the full data, the following harmonic mean estimator of the likelihood of the ith subset is a very efficient estimator given that we do not need to create new MCMC samples based on subsets of the data. Suppose $\boldsymbol{\beta}^{(t)}$ is the sample for the unknown parameter $\boldsymbol{\beta}$ for a model M from the tth iteration of a single MCMC chain for the posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{y})$, then we propose the MC estimator for Eq. (3.3):

$$\hat{p}_i = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(\boldsymbol{y}_{s_i}|\boldsymbol{\beta}^{(t)})}\right)^{-1}.$$

4. SIMULATION EXAMPLE

We consider a small example to compare the different measures in model selection. We generate X_1, X_2, X_3, X_4, X_5, and ε each independently from a standard normal $N(0, 1)$ distribution. Further, we let $X_6 = X_1^2$, $X_7 = X_1 * X_2$, $X_8 = X_3^2$, $X_9 = X_2 * X_3$ and compute $Y = 2 + X_1 + 2X_2 + 3X_3 + X_7 + X_8 + \epsilon$. We repeat this for 500 times to obtain our data set. Then we fit the data set with the following five linear models, each with an intercept: M_1 with X_1, X_2, X_3, X_7 and X_8, M_2 with X_1, X_2 and X_3, M_3 with X_1, X_2, X_4, X_6 and X_7, M_4 with X_1, X_2, X_3, X_8 and X_9 as predictors, and M_5 includes X_2, X_3, X_4, X_5, X_8, X_9 as predictors. We use normal prior $\pi(\beta_i) \overset{iid}{\sim} N(0, 1)$ for the linear coefficients and AIC, DIC, LPML, WAIC1, WAIC2, and BDC as the model selection criteria. For BDC, we use $\boldsymbol{y}_{s_i} = y_i$ and $\boldsymbol{y}_{-s_i} = \boldsymbol{y}\backslash y_i$, which is known as the leave-one-out cross

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

validation approach. For different ψ functions, we choose

$$\psi_1(\boldsymbol{p}) = \sum_{i=1}^{n} p_i \log p_i,$$

$$\psi_2(\boldsymbol{p}) = -\sum_{i=1}^{n} \log p_i,$$

$$\psi_3(\boldsymbol{p}) = ||\boldsymbol{p}||^2 = \sum_{i=1}^{n} p_i^2.$$

Then the corresponding BDs are:

$$BD_{\psi_1}(\bar{\boldsymbol{p}}, \boldsymbol{p}^{(M^k)}) = \sum_{i=1}^{n} \left(\bar{p}_i \log \frac{\bar{p}_i}{p_i^{(M^k)}} - \bar{p}_i + p_i^{(M^k)} \right),$$

$$BD_{\psi_2}(\bar{\boldsymbol{p}}, \boldsymbol{p}^{(M^k)}) = \sum_{i=1}^{n} \left(\frac{\bar{p}_i}{p_i^{(M^k)}} - \log \frac{\bar{p}_i}{p_i^{(M^k)}} - 1 \right),$$

$$BD_{\psi_3}(\bar{\boldsymbol{p}}, \boldsymbol{p}^{(M^k)}) = \sum_{i=1}^{n} (\bar{p}_i - p_i^{(M^k)})^2.$$

WAIC (Watanabe [30]) is a more fully Bayesian approach for estimating the out-of-sample expectation, starting with the computed log point-wise posterior predictive density and then adding a correction for the effective number of parameters to adjust for over-fitting. Two adjustments (Gelman, Hwang and Vehtari [10]) have been proposed in the literature. Both are based on point-wise calculations and can be viewed as approximations to cross-validation. The first approach uses the difference:

$$p_{WAIC1} = 2 \sum_i (\log(E_{\text{post}} p(y_i|\theta)) - E_{\text{post}}(\log p(y_i|\theta))).$$

The other approach uses the variance of individual terms in the log predictive density:

$$p_{WAIC2} = \sum_i \text{Var}_{\text{post}}(\log p(y_i|\theta)).$$

We use the R package "blmeco" to calculate two types of WAIC for each model.

The LPML is obtained by summing the log CPO for all sets:

$$\text{LPML} = \sum_i \log \text{CPO}_i.$$

Table 1 displays the goodness-of-fit scores using different criteria.

We can see from this small example, all the criteria select the same model M_1, which is the correct model.

5. REAL DATA ANALYSIS

5.1. Illustrative Example with Exponential, Weibull, and Semiparametric Cox Models

We study the survival times of kidney transplant patients and their risk factors. The 863 patients were in The Ohio State University Transplant Center from 1982 to 1992 [19]. The

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

TABLE 1. Goodness-of-Fit Measures Based on the Simulated Data Set

Model	AIC	DIC	LPML	WAIC1	WAIC2	BDC_{ψ_1}	BDC_{ψ_2}	BDC_{ψ_3}
1	1400.1	1398.8	−699.8	1397.7	1398.1	0	0	0
2	2090.1	2802.8	−1411.7	2091.4	2091.5	139.2	2.5×10^{20}	15.9
3	2659.0	6737.4	−3403.9	2656.1	2656.6	684.9	5.5×10^{46}	30.4
4	1753.1	1879.6	−945.7	1753.5	1754.0	45.6	8.2×10^{5}	7.2
5	1970.4	2392.0	−1204.0	1968.3	1968.8	99.4	7.4×10^{10}	13.3

TABLE 2. Model Selection Results for the Exponential Models

Model	LPML	DIC	BDC_{ψ_1}	BDC_{ψ_2}	BDC_{ψ_3}
1	−714.0	1429.7	3.2	49.4	4.2
2	−697.9	1405.2	0.4	2.8	0.5
3	−704.1	1412.0	1.6	12.8	2.0
4	−695.3	1391.7	5.6×10^{-3}	0.03	0.01

TABLE 3. Model Selection Results for the Weibull Models

Model	LPML	DIC	BDC_{ψ_1}	BDC_{ψ_2}	BDC_{ψ_3}
1	−1424.9	1187.1	0	0	0
2	−1493.2	1135.1	5.0	37.4	7.2
3	−1467.3	1134.7	5.6	19.3	7.5
4	−1455.1	1135.1	1.2	9.8	1.8

data set contains: the time to death (Y), each patient's gender (X_1), race (X_2), age (X_3), indicator of whether the patient is dead or not (V). We can construct the exponential model and the Weibull model to fit the data, and use the BD to select the model.

First, we apply the exponential model. Based on the previous discussion, we have

$$L(\boldsymbol{\beta}|D) = \exp\left\{\sum_i (v_i Z_i'\boldsymbol{\beta} - Y_i \exp(Z_i'\boldsymbol{\beta}))\right\}.$$

By applying different matrix Z, we can have different models. Let us define $M_1 : Z = (X_1, X_2); M_2 : Z = (X_1, X_3); M_3 : Z = (X_2, X_3); M_4 : Z = (X_1, X_2, X_3)$ as possible models. And we use the same BDs as in the previous example.

We also fit the data using the Weibull model:

$$L(\boldsymbol{\beta}|D) = \alpha^{\sum_i v_i} \exp\left\{\sum_i (v_i Z_i'\boldsymbol{\beta} + v_i(\alpha-1)\log(Y_i) - \exp(Z_i'\boldsymbol{\beta})Y_i^{\alpha})\right\}.$$

The priors of the linear coefficients are independent standard normal distributions in both models and the prior of α in the Weibull model is $\mathcal{G}(1, 1)$.

Tables 2 and 3 exhibit the posterior risk for each of the four models M_1 to M_4 under different criteria for exponential and Weibull models, respectively.

While all criteria agree on model 4 as the best model for the exponential models, LPML and BDCs select model 1 but DIC prefers model 3 while fitting with the family of Weibull

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

TABLE 4. Criteria Results for Semi-Parametric Cox Model

Model	LPML	DIC	BDC_{ψ_1}	BDC_{ψ_2}	BDC_{ψ_3}
1	−2670.9	4324.4	0	0	0
2	−3242.1	4281.4	29.7	6.7×10^8	6.9
3	−3554.3	4280.8	34.0	4.1×10^{12}	3.4
4	−3339.6	4281.9	24.2	8.4×10^{19}	3.8

TABLE 5. Summary Statistics for the Posterior Absolute Values of the Regression Coefficients

Min	$Q1$	Median	$Q3$	Max
2.279×10^{-4}	0.345	0.488	0.628	2.039

models. For the Weibull models, each criterion prefers a smaller model with only gender and race as predictors, except DIC selects race and age as predictors.

Then we fit the semi-parametric Cox model with a gamma process prior. Here we select $K = 10$ sub-intervals and the gamma process prior on the piecewise constant hazard function with $c_0 = 2$ in Eq. (2.2). The reason for choosing $c_0 = 2$ is to obtain a reasonably fast convergence and good performance of isolating the true non-zero predictors. We set $\alpha_{0j} = c_0 H^*(s_j) = c_0 \gamma_0 s_j$ and set γ_0 to be 1. Table 4 displays the posterior risk for the BDC measure with the three convex functions ψ_1, ψ_2, and ψ_3 along with other criteria DIC and LPML. The conclusion on model selection is the same as that of the Weibull models.

Let us observe in both Weibull and semi-parametric examples DIC produces different results from other criteria. This can be explained by their different approaches; DIC is a model selection criterion, which is based on the whole data, LPML, and BDCs, use predictive approaches, which depend on leave-one-out cross-validated sub-data sets. Moreover, the prior and outliers (even the same for DIC and other criteria) have different effects on model selection due to the two different approaches.

5.2. An Example of Large p Small n

In this real data example, we have analyzed an ovarian cancer gene expression data set [15,31]. The expressions of a list of 2647 genes and the survival/censoring times of 153 ovarian cancer patients treated at the Memorial Sloan Kettering cancer center are used, where 24 of the 153 subjects are censored. We fit an exponential survival model. Since it is a "large p, small n" problem, we use the Bayesian elastic net method. We use R-package "psbcGroup" [23] to perform the Bayesian elastic net calculation. Table 5 and Figure 2 summarize the absolute values of the posterior mean of all coefficients. With different cut-off points for the regression coefficients, we have different number of genes included. Table 6 summarizes the number of genes selected for each model, where the cut-off points are selected to be 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.6, and 1.7. We then apply different criteria to select the best cut-off point. Table 7 summarizes the posterior risk for each of the eight models using three different convex functions ψ_1, ψ_2, or ψ_3 along with DIC and LPML.

From Table 7 we can see all criteria select model 7, which uses 1.6 as the cut-off value and contains 3 genes. Too few genes selected fails to provide accurate predictions, while too many genes (>3) selected also increases the "prediction" error.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

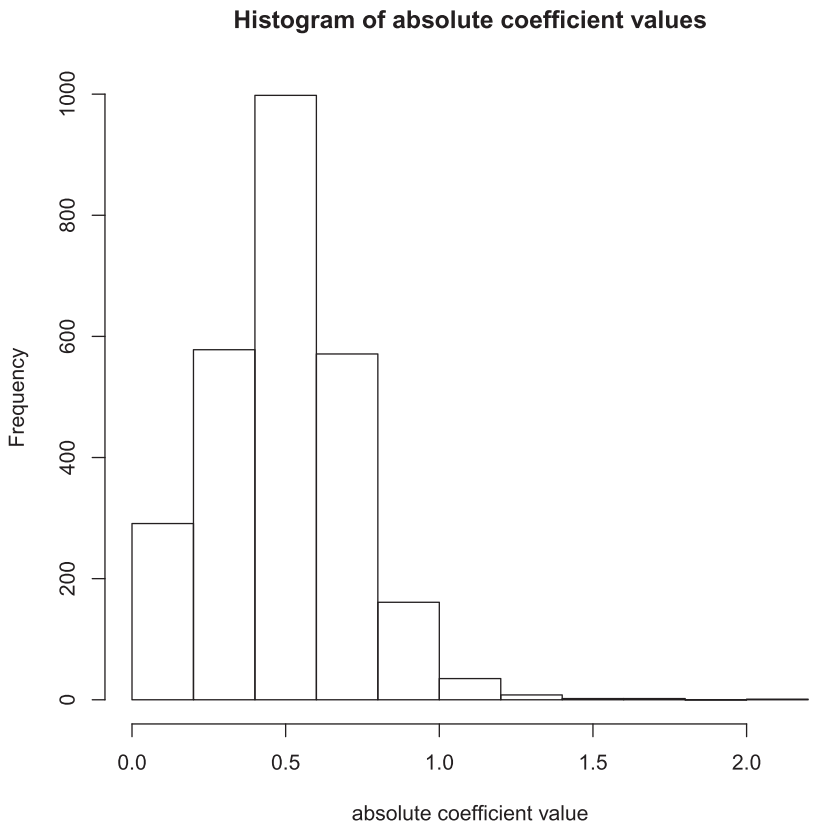


FIGURE **2.** Histogram of absolute regression coefficients.

TABLE 6. Number of Genes Selected in Sparse Models with Specified Cut-Off Values

Model	1	2	3	4	5	6	7	8
Cut-off Value	0.9	1	1.1	1.2	1.3	1.4	1.6	1.7
Number of selected genes	95	48	22	13	8	5	3	1

TABLE 7. Model Selection Results

Model	LPML	DIC	BDC_{ψ_1}	BDC_{ψ_2}	BDC_{ψ_3}
1	−1037.6	1319.4	9.2	3.4×10^{13}	7.4
2	−832.2	1298.3	13.7	3.9×10^{17}	5.7
3	−729.3	1279.1	4.8	2.1×10^{3}	2.7
4	−664.0	1280.6	1.4	86.9	0.5
5	−633.8	1283.0	0.6	38.8	0.2
6	−639.6	1274.3	5.3×10^{-2}	2.9	2.6×10^{-2}
7	−635.8	1271.8	3.5×10^{-4}	8.3×10^{-3}	1.3×10^{-4}
8	−638.2	1278.3	3.2×10^{-1}	4.5	0.1

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

6. DISCUSSION

Unlike LPML, BDC provides a more flexible way to build a criterion through the predictive probabilities. Different choices of subset further improve the flexibility. Although BDC gives us a general form to select models, the choice of the convex function $\psi(x)$ usually depends on the utility function of the users. For example, if overestimation is a more serious mistake than underestimation, then we should choose a linex loss function for the ψ. If the users do not know how to specify their utility functions, we can empirically evaluate the posterior risks for several convex loss functions and select the function with the smallest posterior risk. For example, Table 7 shows among ψ_1 to ψ_3, ψ_3 function (the Euclidean distance function) is preferred for this data set, for models 1 to 5 and ψ_1 is preferred for models 6 to 8 because it yields the smallest prediction errors among the three convex loss functions. Our simulation and real data analysis also suggest that the choice of the best model is relatively insensitive to the choice of the ψ function. Further investigation is needed to follow up on this suggestion.

References

1. Banerjee, A., Merugu, S., Dhillon, I.S., & Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research* 6: 1705–1749.
2. Berger, J.O. & Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* 91(433): 109–122.
3. Bernardo, J.M. & Smith, A.F. (2000). Bayesian Theory. Somerset, NJ: John Wiley & Sons.
4. Bregman, L.M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* 7(3): 200–217.
5. Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)* 34(2): 187–220.
6. Cox, D.R. (1980). Local ancillarity. *Biometrika* 67(2): 279–286.
7. Dawid, A.P. (1984). Present position and potential developments: some personal views: statistical theory: the prequential approach. *Journal of the Royal Statistical Society. Series A (General)* 147(2): 278–292.
8. Geisser, S. & Eddy, W.F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* 74(365): 153–160.
9. Gelfand, A.E. & Dey, D.K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B* 56(3): 501–514.
10. Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics And Computing* 24(6): 997–1016.
11. Goh, G. (2015). Applications of Bregman divergence measures in Bayesian modeling. Doctoral Dissertations, 785, University of Connecticut, Storrs, CT, USA. http://digitalcommons.uconn.edu/dissertations/785.
12. Goh, G. & Dey, K.D. (2017) Bayesian model assessment and selection using Bregman divergence. Preprint. Bayesian Analysis, under review.
13. Guerrier, S. & Victoria-Feser, M. (2015) A prediction divergence criterion for model selection. arXiv: 1511.04485v1.
14. Gui, J. & Li, H. (2005). Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* 21(13): 3001–3008.
15. Higgins, M.E., Claremont, M., Major, J.E., Sander, C., & Lash, A.E. (2007). CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Research* 35(suppl 1): D721–D726.
16. Ibrahim, J.G., Chen, M.H., Zhang, D., & Sinha, D. (2013). Bayesian analysis of the Cox model. *Handbook of Survival Analysis* 27–48.
17. Itakura, F. & Saito, S. (1970). A statistical method for estimation of speech spectral density and formant frequencies. *Electronics and Comunications in Japan* 53: 36–43.
18. Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)* 40(2): 214–221.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press

19. Klein, J.P. & Moeschberger, M.L. (2003). Survival Analysis: Techniques for Censored and Truncated Data. New York, NY: Springer Science & Business Media.
20. Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian LASSOs. *Bayesian Analysis* 5(2): 369–411.
21. Lee, K.H., Chakraborty, S., & Sun, J. (2011). Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. *The International Journal of Biostatistics* 7(1): 1–32.
22. Lee, K.H., Chakraborty, S., & Sun, J. (2015). Survival prediction and variable selection with simultaneous shrinkage and grouping priors. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8(2): 114–127.
23. Lee, K.H., Chakraborty, S., & Sun, J. (2017). Penalized Parametric and Semiparametric Bayesian Survival Models with Shrinkage and Grouping Priors. R-Package https://cran.r-project.org/web/packages/psbcGroup/psbcGroup.pdf.
24. Lewis, P.O., Xie, W., Chen, M.H., Fan, Y., & Kuo, L. (2014). Posterior predictive Bayesian phylogenetic model selectoion. *Systematic Biology* 63(3): 309–321.
25. Li, Q. & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis* 5(1): 151–170.
26. Park, T. & Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association* 103(482): 681–686.
27. Spiegelhalter, D., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* 64(4): 583–639.
28. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36(2): 111–147.
29. Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* 16(4): 385–395.
30. Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11: 3571–3594.
31. Zhang, W., Ota, T., Shridhar, V., Chien, J., Wu, B., & Kuang, R. (2013). Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Computational Biology* 9(3): e1002975.
32. Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2): 301–320.

https://doi.org/10.1017/S0269964818000190 Published online by Cambridge University Press