Statistica Sinica **11**(2001), 419-443

CRITERION-BASED METHODS FOR BAYESIAN MODEL ASSESSMENT

Joseph G. Ibrahim, Ming-Hui Chen and Debajyoti Sinha

Harvard School of Public Health and Dana-Farber Cancer Institute, Worcester Polytechnic Institute and Medical University of South Carolina

Abstract: We propose a general Bayesian criterion for model assessment. The criterion is constructed from the posterior predictive distribution of the data, and can be written as a sum of two components, one involving the means of the posterior predictive distribution and the other involving the variances. It can be viewed as a Bayesian goodness-of-fit statistic which measures the performance of a model by a combination of how close its predictions are to the observed data and the variability of the predictions. We call this proposed predictive criterion the *L measure*, it is motivated by earlier work of Ibrahim and Laud (1994) and related to a criterion of Gelfand and Ghosh (1998). We examine the L measure in detail for the class of generalized linear models and survival models with right censored or interval censored data. We also propose a calibration of the L measure, defined as the prior predictive distribution of the difference between the L measures of the candidate model and the criterion minimizing model, and call it the *calibration distribution*. The calibration distribution will allow us to formally compare two competing models based on their L measure values. We discuss theoretical properties of the calibration distribution in detail, and provide Monte Carlo methods for computing it. For the linear model, we derive an analytic closed form expression for the L measure and the calibration distribution, and also derive a closed form expression for the mean of the calibration distribution. These novel developments will enable us to fully characterize the properties of the L measure for each model under consideration and will facilitate a direct formal comparison between several models, including non-nested models. Informative priors based on historical data and computational techniques are discussed. Several simulated and real datasets are used to demonstrate the proposed methodology.

Key words and phrases: Calibration, model selection, predictive criterion, predictive distribution, variable selection.

1. Introduction

Model assessment and model comparison is a crucial part of statistical analysis. Due to recent computational advances, sophisticated techniques for Bayesian model assessment are becoming increasingly popular. We have seen a recent surge in the statistical literature on Bayesian methods for model assessment and model comparison, including George and McCulloch (1993), Geisser (1993),

Madigan and Raftery (1994), Ibrahim and Laud (1994), Bernardo and Smith (1994), Laud and Ibrahim (1995), Kass and Raftery (1995), Raftery, Madigan and Volinsky (1995), George, McCulloch and Tsay (1996), Gelman, Meng and Stern (1996), Raftery, Madigan and Hoeting (1997), Gelfand and Ghosh (1998), Clyde (1999) and Spiegelhalter, Best and Carlin (1998). The scope of Bayesian model assessment is quite broad, and can be investigated via model diagnostics, goodness-of-fit measures, or posterior model probabilities (or Bayes factors).

Many of the proposed Bayesian methods for model comparison rely on posterior model probabilities or Bayes factors, and it is well known that to use these methods, proper prior distributions are needed. It is usually a major task to specify prior distributions for all models under consideration, especially if the model space is large. This issue has been discussed in detail by several authors, including Ibrahim and Laud (1994), Laud and Ibrahim (1995) and Chen, Ibrahim and Yiannoutsos (1999). In addition, it is well known that posterior model probabilities are generally sensitive to the choices of prior parameters, and thus one cannot simply select vague proper priors to get around the elicitation issue. Thus, computing posterior model probabilities can become a monumental chore if informative prior distributions are difficult to specify. Alternatively, criterion-based methods can be attractive in the sense that they do not require proper prior distributions in general, and thus have an advantage over posterior model probabilities in this sense. Several recent papers advocating the use of Bayesian criteria for model assessment include Ibrahim and Laud (1994), Laud and Ibrahim (1995), Gelman, Meng and Stern (1996), Gelfand and Ghosh (1998), and Spiegelhalter, Best and Carlin (1998). However, posterior model probabilities are intrinsically well-calibrated since probabilities are relatively easy to interpret, whereas criterion-based methods are generally not easy to calibrate or interpret. Thus, one potential criticism of criterion-based methods for model comparison is that they generally do not have well defined calibrations. Thus, one needs to calibrate these criteria somehow so that they can be more easily interpreted.

To motivate the notion of calibration for our proposed methodology, we consider the landmark AIDS study, ACTG036. The ACTG036 study was a double blind placebo-controlled clinical trial comparing zidovudine (AZT) to placebo in persons with CD4 counts less than 500. The sample size for this study, excluding cases with missing data, was $n = 183$. The response variable (y) for these data is binary with a 1 indicating death, development of AIDS, or AIDS related complex (ARC), and a 0 indicates otherwise. Several covariates were also measured. The ones we use here are CD4 count (x_1) (cell count per mm^3 of serum), age (x_2), treatment, (x_3), and race (x_4). The covariates CD4 count and age are continuous, whereas the other covariates are binary. We consider a logistic regression for the response y as a function of the covariates (see Chen, Ibrahim and Yiannoutsos

(1999)). For the ACTG036 data, suppose we consider the two competing models x_1 and the full model (x_1, x_2, x_3, x_4), which is also the criterion minimizing model with respect to the L measure. The L measure values for these models are 16.82 and 16.30, respectively (See Table 3). Now we ask the question, how "far" is model (x_1) from the criterion-minimizing model (x_1, x_2, x_3, x_4)? Can we formally quantify the size difference in the criterion values between the two models, that is, can we calibrate the L measure? These are the types of questions we wish to address. This example thus serves as a good demonstration of the need for calibrated model selection criteria for comparing models.

In this article, we propose a Bayesian criterion, called the *L measure*, for model assessment and model comparison, and also propose a calibration for it. The L measure criterion is constructed from the posterior predictive distribution of the data, and can be written as a sum of two components, one involving the means of the posterior predictive distribution and the other involving the variances. It can be viewed as a statistic which measures the performance of a model by a combination of how close its predictions are to the observed data, and the variability of the predictions. We argue that the proposed criterion can be used as a general model assessment tool for comparing models and assessing goodness-of-fit for a particular model, and thus is potentially quite versatile. To facilitate the formal comparison of several models, we propose a novel calibration for the L measure by deriving the marginal prior predictive density of the difference between the L measures of the candidate model and the true model. We call the calibrating marginal density the *calibration distribution*. Since, in practice, the true model will not be known, we use the criterion minimizing model in place of the true model, and give its theoretical justification. Thus an L measure statistic and its corresponding calibration distribution are computed for each candidate model. We discuss theoretical properties of the calibration distribution in detail. For the linear model, we derive an analytic closed form expression for the L measure and the calibration distribution. Also, we derive an analytic expression for the mean of the calibration distribution, which will serve as a useful summary measure. Although the L measure and calibration distribution are quite general in their definition and could be applied to any parametric or semi-parametric model, we focus our applications on generalized linear models (GLM's) and survival models with right censored or interval censored data.

The rest of this paper is organized as follows. In Section 2, we motivate and derive the L measure. In Sections 2.1 and 2.2 we discuss the L measure for GLM's and models for survival data, and we present computational formulas and simplified expressions for the criterion. We also provide Monte Carlo (MC) methods for computing the L measure. In Section 3, we motivate the calibration of the L measure, derive the calibration distribution, derive its properties, and

give MC strategies for computing it. In Section 4, we discuss how informative prior elicitation based on historical data can facilitate the computation of the calibration distribution. In Section 5, several theoretical properties of the L measure and calibration distribution are derived for the linear model. Specifically, we obtain analytic expressions for the L measure and calibration distribution, as well as obtaining analytic expressions for the mean of the calibration distribution. Several theorems are presented which characterize properties of the L measure and calibration distribution under various settings involving nested and non-nested models. In Section 6, we present several examples with real data sets that demonstrate the methodology.

2. The L Measure

Consider an experiment that yields the data $y = (y_1, \ldots, y_n)$. Denote the joint sampling density of the y_i's by $p(y|\theta)$, where θ is a vector of indexing parameters. In this article, we allow the y_i's to be fully observed, right censored, or interval censored. In the right censored case, y_i may be a failure time or a censored time. In the interval censored case, we only observe the interval $[a_{l_i}, a_{r_i}]$ in which y_i occurred. Let $z = (z_1, \ldots, z_n)$ denote future values of an imagined replicate experiment. That is, z is a future response vector with the same sampling density as $y|\theta$. The idea of using a future response vector z in developing a criterion for assessing a model or comparing several models has been well motivated in the literature by Geisser (1993) and the many references therein, Ibrahim and Laud (1994), Laud and Ibrahim (1995), and Gelfand and Ghosh (1998). The imagined replicate experiment makes y and z directly comparable, and exchangeable a priori. It seems clear that good models should make predictions close to what has been observed for an identical experiment. With this notion in mind, Ibrahim and Laud (1994) defined their statistic as the expected squared Euclidean distance between y and z,

$$L_{IL} = E[(z-y)'(z-y)] \ , \tag{2.1}$$

where the expectation is taken with respect to the posterior predictive distribution of $z|y$,

$$p(z|y) = \int p(z|\theta)\ p(\theta|y)\ d\theta \ . \tag{2.2}$$

Here θ denotes the vector of indexing parameters, $p(z|\theta)$ is the sampling distribution of the future vector z, and $p(\theta|y)$ denotes the posterior distribution of θ. Straightforward algebra shows that L_{IL} can be written as

$$L_{IL} = \sum_{i=1}^{n} \Big(\mathrm{Var}(z_i|y) + (E(z_i|y) - y_i)^2 \Big) \ , \tag{2.3}$$

and thus L_{IL} can be written as a sum of two terms, one involving the predictive variances and the other term is like a bias term involving the squared difference between the predictive means and the observed data.

A more general version of (2.1) is as follows. For a given model, we first define the statistic

$$L_1(y, b, k) = E\left[(z-b)'(z-b)\right] + k(y-b)'(y-b)\ , \tag{2.4}$$

where the expectation is taken with respect to the posterior predictive distribution of $z|y$. We note here that y and z may represent a transformation of the original data. For example, in survival analysis, it is common to take the logarithms of the survival times, and in this case y would represent the logs of the survival times. Logarithms are also a common transformation in Poisson regression.

The statistic in (2.4) takes the form of a weighted discrepancy measure. The vector $b = (b_1, \ldots, b_n)$ is an arbitrary location vector to be chosen, and k is a nonnegative scalar that weights the discrepancy based on the future values relative to the observed data. Setting $b = y$, we get the criterion of Ibrahim and Laud (1994) given in (2.1). The case $k = 0$ can be interpreted as a squared error loss discrepancy measure in the future response vector. The general criterion in (2.4) is a special case of a class considered by Gelfand and Ghosh (1998), motivated by a Bayesian decision theoretic viewpoint. We refer the reader to their paper for more detail.

In scalar notation, (2.4) can be written as

$$L_1(y, b, k) = \sum_{i=1}^{n} \left\{ Var(z_i|y) + (\mu_i - b_i)^2 + k(y_i - b_i)^2 \right\}, \tag{2.5}$$

where $\mu_i = E(z_i|y)$. We follow Gelfand and Ghosh (1998) by selecting b as the minimizer of (2.5),

$$\hat{b} = (1-\nu)\mu_i + \nu\ y_i, \tag{2.6}$$

where $\nu = k/(k+1)$. Upon substitution in (2.5),

$$L_2(y, \nu) = \sum_{i=1}^{n} Var(z_i|y_i) + \nu \sum_{i=1}^{n} (\mu_i - y_i)^2. \tag{2.7}$$

Clearly $0 < \nu < 1$, where $\nu = 0$ if $k = 0$, and $\nu \to 1$ as $k \to \infty$. The quantity ν plays a major role in (2.7) (see Section 5). Ibrahim and Laud (1994) use $\nu = 1$, and thus give equal weight to the squared bias and variance components. However there is no theoretical justification for such a weight and, indeed, using $\nu = 1$ may not be desirable in certain situations. Allowing ν to vary between

zero and one gives the user a great deal of flexibility in the tradeoff between bias and variance. This suggests the question of whether certain values of ν are "optimal" in some sense for model selection purposes. In Section 5, we address this optimality issue for the linear model and show that certain values of ν yield desirable properties of the L measure and the calibration distribution. These are novel developments and appear to shed light on the role and interpretation of ν in model selection.

If y is fully observed, (2.7) is straightforward to compute. However if y contains right censored or interval censored observations, then (2.7) is computed by taking the expectation of these censored observations with respect to this posterior predictive distribution. Let $y = (y_{obs}, y_{cens})$, where y_{obs} denotes the completely observed components of y, and y_{cens} denotes the censored components. Here we assume that y_{cens} is a random quantity and $a_l < y_{cens} < a_r$, where a_l and a_r are known. For ease of exposition, we let $D = (n, y_{obs}, a_l, a_r)$ denote the observed data. Then (2.7) is modified as

$$L(y_{obs}, \nu) = E_{y_{cens}|D}[I(a_l < y_{cens} < a_r) L_2(y, \nu)], \tag{2.8}$$

where $I(a_l < y_{cens} < a_r)$ is 1 if $a_l < y_{cens} < a_r$ and 0 otherwise. (Note that when a_l, y_{cens}, and a_r are vectors, $a_l < y_{cens} < a_r$ means that the double inequalities hold for each component of these vectors). We call (2.8) the *L measure*, small values implying a good model. Specifically, we can write (2.8) as

$$L(y_{obs}, \nu) = \int \int_{a_l}^{a_r} L_2(y, \nu)\ p(y_{cens}|\theta)\ p(\theta|D)\ dy_{cens}\ d\theta, \tag{2.9}$$

where $p(y_{cens}|\theta)$ is the sampling density of y_{cens} and $p(\theta|D)$ is the posterior density of θ given the observed data D. If y has right censored observations, $a_r = \infty$ and a_l is a vector of censoring times; if y has interval censored observations, (a_l, a_r) is a sequence of finite interval censoring times; if y is fully observed, (2.8) reduces to (2.7) and $L(y_{obs}, \nu) \equiv L_2(y, \nu)$. We use (2.8) throughout as our Bayesian criterion for model assessment and model comparison. We note here that our definition in (2.8) is quite different from the development of Gelfand and Ghosh (1998) and Sinha, Chen and Ghosh (1999).

It can be shown that

$$L_2(y, \nu) = \sum_{i=1}^{n} \left\{ E_{\theta|D} \left(E\left[(z_i)^2 | \theta \right] \right) - \mu_i^2 \right\} + \nu \sum_{i=1}^{n} (\mu_i - y_i)^2, \tag{2.10}$$

where $\mu_i = E_{\theta|D}[E(z_i|\theta)]$. Thus (2.7) and (2.8) can be computed by sampling from the posterior distribution of θ. Once the posterior samples of θ are obtained, (2.10) and (2.8) can be evaluated. More specifically, suppose that

$\{\theta^{(r)},\ r = 1, 2, \ldots, R\}$ is a sample from $p(\theta|D)$ and $\{y_{cens}^{(r)},\ r = 1, 2, \ldots, R\}$ is a sample from the truncated posterior predictive distribution $I(a_l < y_{cens} < a_r)\ p(y_{cens}|\theta)\ p(\theta|D)$. Then, an MC estimate of $L(y_{obs}, \nu)$ is given by

$$\hat{L}(y_{obs}, \nu) = \sum_{i=1}^{n} \left\{ \frac{1}{R} \sum_{r=1}^{R} \left(E\left[(z_i)^2 | \theta^{(r)} \right] \right) - \hat{\mu}_i^2 \right\}$$
$$+\nu \left\{ \sum_{i:\ y_i \text{ observed}} (\hat{\mu}_i - y_i)^2 + \frac{1}{R} \sum_{r=1}^{R} \left[\sum_{i:\ y_i \text{ censored}} (\hat{\mu}_i - y_{cens,i}^{(r)})^2 \right] \right\}, \quad (2.11)$$

where $\hat{\mu}_i = \dfrac{1}{R} \sum_{r=1}^{R} E(z_i|\theta^{(r)})$. If $E\left[(z_i)^2|\theta\right]$ and $E(z_i|\theta)$ are not analytically available, we need an MC sample $\{z^{(r)},\ r = 1, 2, \ldots, R\}$ from the complete posterior predictive distribution $p(z|\theta)p(\theta|D)$. Then, in (2.11), we replace

$$\frac{1}{R} \sum_{r=1}^{R} (E[(z_i)^2|\theta^{(r)}]) \text{ and } \frac{1}{R} \sum_{r=1}^{R} E(z_i|\theta^{(r)}) \text{ by } \frac{1}{R} \sum_{r=1}^{R} (z_i^{(r)})^2 \text{ and} \frac{1}{R} \sum_{r=1}^{R} z_i^{(r)}.$$

Thus, computing $L(y_{obs}, \nu)$ is relatively straightforward, but may be intensive for some models. Gibbs sampling from the posterior distribution of θ for generalized linear models and most survival data models is straightforward.

If the posterior expectation in (2.10) is finite, $L_2(y, \nu)$ is well defined. For the class of GLM's and the survival models we consider here, the posterior expectation in (2.10) exists when the data is left untransformed or is transformed to logarithms. The posterior expectation in (2.10) can exist even when the prior for θ is improper.

2.1. L measure for generalized linear models

Suppose that $y_1, \ldots, y_n$ are independent observations with density

$$p(y_i \mid x_i, \theta_i, \tau) = \exp\left\{ a_i^{-1}(\tau)(y_i\theta_i - b(\theta_i)) + c(y_i, \tau) \right\}\ ,\ i = 1, \ldots, n\ , \quad (2.12)$$

indexed by canonical parameter θ_i and scale parameter τ, where $b(.)$ $c(.)$ are known functions. Further suppose the θ_i's satisfy the equations

$$\theta_i = \theta(\eta_i)\ ,\ i = 1, \ldots, n\ , \quad (2.13)$$

$$\eta = X\beta\ , \quad (2.14)$$

where $\eta_i = x_i'\beta$ are the components of η, X is an $n \times p$ full rank matrix of covariates, $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, and θ is a monotonic differentiable function. We assume that the covariates are fixed

throughout and, for ease of exposition, we assume $a_i^{-1}(\tau) = 1$, as in logistic and Poisson regression.

For the class of generalized linear models (GLM's), $L_2(y, \nu) \equiv L(y_{obs}, \nu)$ and we can write

$$L(y_{obs}, \nu) = \sum_{i=1}^{n} \left[E_{\beta|y}(b''(\theta_i)) - \left\{ E_{\beta|y}(b'(\theta_i)) \right\}^2 \right] + \nu \sum_{i=1}^{n} \left[E_{\beta|y} \left(b'(\theta_i) \right) - y_i \right]^2 , \tag{2.15}$$

where primes indicate differentiation with respect to θ_i. We mention here that $b'(.)$ and $b''(.)$ are the mean and variance functions of the GLM, respectively, and have simple closed forms as functions of β (see McCullagh and Nelder (1989)). Thus, $L(y_{obs}, \nu)$ is easily computed for GLM's via Gibbs sampling, once samples from the posterior distribution of $\beta|y$ are obtained.

2.2. L measure for survival data

A proportional hazards model is defined by a hazard function of the form

$$h(y, x) = h_0(y) \exp(x'\beta) , \tag{2.16}$$

where $h_0(y)$ denotes the baseline hazard function at time y, x denotes the covariate vector for an arbitrary individual in the population, and β denotes a vector of regression coefficients. The likelihood function for a set of right censored data on n individuals in a proportional hazards model is

$$L\left(\beta, h_0(y)\right) = \prod_{i=1}^{n} \left[h_0(y_i) \exp(\eta_i) \right]^{\delta_i} \left(S_0(y_i)^{\exp(\eta_i)} \right), \tag{2.17}$$

where $\eta_i = x_i'\beta$, y_i is an observed failure time or censoring time for the ith individual and δ_i is 1 if y_i is a failure time and 0 if it is a censoring time. Further, $S_0(\cdot)$ is the baseline survivor function which, since we consider continuous survival distributions, is related to $h_0(\cdot)$ by $S_0(y) = \exp\left(-\int_0^y h_0(u)\, du\right)$. The log transformation of survival time is quite common so as to define the observables (survival times) on the entire real line, and often results in a more symmetric posterior predictive distribution for $z|y$.

For the exponential model with right censored data, we have $h_0(y) = 1$ and

$$L(y_{obs}, \nu)) = \int L_2(y, \nu)\ p(y_{cens}|D)\ dy_{cens}, \tag{2.18}$$

where

$$L_2(y, \nu) = \sum_{i=1}^{n} \left[E_{\beta|D}(\exp(-2\eta_i)) - \left\{ E_{\beta|D}(\exp(-\eta_i)) \right\}^2 \right] + \nu \sum_{i=1}^{n} \left[E_{\beta|D}(\exp(-\eta_i)) - y_i \right]^2 .$$

Similar expressions for $L_2(y, \nu)$ can be obtained for the Weibull or extreme value models but without closed forms for $L_2(y, \nu)$. For example, the L measure can be computed for semi-parametric survival models in which a suitable prior process is defined for the unspecified baseline hazard rate. See Sinha and Dey (1997) for a recent review of semiparametric Bayesian survival analysis and prior processes for hazard rates. We can also entertain semi-parametric survival models with interval censored survival data as well as non-proportional hazards models, models with time varying covariates, and frailty models. We demonstrate the L measure for parametric and semi-parametric survival models with right censored and interval censored data in Section 6. In addition, we examine survival models with time varying covariates.

3. The Calibration Distribution

Criterion-based methods typically rely on the minimum criterion value as the basis for model selection. However this basis is not satisfactory in general, since it does not allow a formal comparison of criterion values between two or more competing models. For example, in variable subset selection, a model with 5 predictors might achieve the minimum criterion value while a model with 3 predictors achieves a slightly larger value. Which model should be chosen? This situation arises often in practice, and in these cases it is desirable to have a calibration of the criterion to formally compare criterion values between the candidate models. Thus, one of the crucial steps in using criterion-based methods for model assessment and model choice is to define a calibration for the criterion.

To motivate our calibration distribution, let $L_c(y_{obs}, \nu)$ denote the L measure for the candidate model c, and let $L_t(y_{obs}, \nu)$ denote the L measure for the true model t. Now consider the difference

$$D(y_{obs}, \nu) \equiv L_c(y_{obs}, \nu) - L_t(y_{obs}, \nu) \ . \tag{3.1}$$

The quantity in (3.1) is a random variable in y_{obs}, and depends on ν. To calibrate the L measure, we construct the *marginal distribution* of $D(y_{obs}, \nu)$, computed with respect to the *prior predictive distribution* of y_{obs} under the true model t:

$$p_t(y_{obs}) = \int p_t(y_{obs}|\theta)\pi_t(\theta) \ d\theta. \tag{3.2}$$

Note that Box (1980) developed similar ideas in calibration using the prior predictive distribution, and Gelman, Meng and Stern (1996) developed calibration measures using the posterior predictive distribution.

Our proposed calibration distribution is

$$p_{L_c} \equiv p(D(y_{obs}, \nu)), \tag{3.3}$$

the marginal distribution of $D(y_{obs}, \nu)$, computed with respect to the prior predictive distribution of y_{obs} under the true model t. We refer to p_{L_c} as the *calibration distribution* for the candidate model c. Since p_{L_c} is a univariate distribution, it is easily tabulated and plotted. If the candidate model is "close" to the true model p_{L_c} should have a mean (or mode) that is close to zero and much of its mass should be centered around this point. On the other hand, if the candidate model and true model are far apart, then p_{L_c} will have a mean (or mode) that is far from zero. Clearly p_{L_c} depends on c. We also see from (3.3) that, for p_{L_c} to be well defined, we need a proper prior distribution for θ. The definition at (3.3) is appealing since it avoids the potential problem of a double use of the data, as discussed by Bayarri and Berger (1999).

The definition of p_{L_c} depends on the data only through y_{obs}. For GLM's, $y_{obs} = y$, and (3.3) is clear. When we have right censored data, y_{obs} consists of observed failure times and p_{L_c} is a function of them and its computation does not depend on censoring times. In situations where all observations are censored as, for example, with case-2 interval censored data (see Section 6), then y_{obs} consists of the empty set. In this case, the definition of p_{L_c} in (3.3) must be slightly modified. We "impute" each interval censored observation by sampling from the truncated prior predictive distribution, where the truncation is taken to be the endpoints of the interval censored observation. Thus, if y_i is interval censored in the interval $[a_{l_i}, a_{r_i}]$, we impute y_i by replacing it with a sample of size 1 from

$$p_t(y_i) \propto \int p_t(y_i|\theta)\ \pi_t(\theta)\ d\theta, \quad a_{l_i} < y_i < a_{r_i}, \tag{3.4}$$

$i = 1, \ldots, n$. We denote the sampled value by $\tilde{y}_i$ for each interval censored observation and take $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)$. We then treat $\tilde{y}$ as y_{obs}. This technique for obtaining the calibration distribution p_{L_c} when all of observations are interval censored produces good results, as demonstrated in Section 6. This technique could also be used when all of the observations in a given data set are right censored, but this scenario is much less common in practice.

Once p_{L_c} is computed, several statistical summaries can be obtained from it. These include various HPD intervals and the mean of $D(y_{obs}, \nu)$. The mean is denoted by

$$\mu_c(\nu) = E_t(D(y_{obs}, \nu))\ , \tag{3.5}$$

where $E_t(.)$ denotes the expectation with respect to the prior predictive distribution of the true model. This summary $\mu_c(\nu)$ is attractive since it measures, on average, how close the centers are of the candidate and true models. We show in Section 5 that $\mu_c(\nu)$ has some attractive properties for the linear model.

Since the true model t will not be known in practice, we use the criterion minimizing model t_{min} in place of t for computing (3.1). Thus, in practice, we compute

$$\hat{D}(y_{obs}, \nu) = L_c(y_{obs}, \nu) - L_{t_{min}}(y_{obs}, \nu) \ , \tag{3.6}$$

$$\hat{p}_{L_c} = p(\hat{D}(y_{obs}, \nu)) \ , \tag{3.7}$$

where $\hat{p}_{L_c}$ is computed with respect to the prior predictive distribution of the criterion minimizing model. Also, $\mu_c(\nu)$ is estimated by $\hat{\mu}(\nu)$, where $\hat{\mu}(\nu) = E_{t_{min}}[\hat{D}(y_{obs}, \nu)]$. It turns out that, for the linear model, the criterion minimizing model yields many of the same theoretical properties as the true model (See Section 5).

We briefly describe how to compute the calibration distribution p_{L_c} via MC sampling. For illustrative purposes, we consider only the case where y_{obs} is *not empty* (the computation is even much simpler when y_{obs} is empty). For a candidate model c,

(i) Generate a pseudo observation $\tilde{y}$ from the prior predictive distribution $p_t(y|\theta)$ $\pi_t(\theta)$.

(ii) Set $y_{obs} = \tilde{y}$ and use the method described in Section 2 to obtain an MC estimate of $L_c(y_{obs}, \nu)$.

We repeat (i) and (ii) R times to obtain an MC sample of $L_c(y_{obs}, \nu)$. Then we repeat (i) and (ii) R times using the criterion minimizing model to obtain an MC sample of $L_{t_{min}}(y_{obs}, \nu)$. Using these MC samples, we can compute the entire calibration distribution p_{L_c}, for example, by using the kernel method (see Silverman (1986)). We note that step (ii) may be computationally intensive but the entire computational procedure is quite straightforward. In the examples given in Section 6, for a medium sample size, say $100 < n < 500$, the entire computational time for obtaining the criterion and the calibration distribution ranges from one to five hours on a standard digital alpha machine.

To compute p_{L_c}, we need proper prior distributions. Although we show in Sections 5 and 6 that p_{L_c} is *not* sensitive to choices of vague proper priors, suitable choices of informative priors can be quite useful in improving the precision in the estimation of p_{L_c}, and hence sharpening the comparisons between the candidate models. Towards this goal, potential choices of informative priors are those developed in Ibrahim, Ryan and Chen (1998), Chen, Ibrahim and Yiannoutsos (1999), and Ibrahim and Chen (1998). These priors are constructed from historical data and appear to be useful for inference and for computing p_{L_c}. We note here that other types of informative priors may be used as well, such as those discussed in Box and Tiao (1973), Gelman et al. (1995), and Carlin and Louis (1996). In the next section, we give a brief review of informative priors based on historical data.

4. Informative Prior Distributions

The informative prior construction is based on the existence of historical data measuring the same response variable and covariates as the current study. We assume only one historical dataset, as the extension to multiple historical datasets is straightforward. Let n_0 denote sample size for the historical data, y_0 be an $n_0 \times 1$ response vector, and X_0 the $n_0 \times p$ matrix of covariates corresponding to y_0. Also, let y_{0i} denote the ith component of y_0, $x_{0i}' = (x_{0i1}, x_{0i2}, \ldots, x_{0ip})$ be the ith row of X_0 with $x_{0i1} = 1$ corresponding to an intercept, and $D_0 = (n_0, y_0, X_0)$ the historical data. Denote the likelihood function of θ based on the historical data D_0 by $L(\theta|D_0)$. The informative prior is

$$\pi(\theta|D_0, a_0) \propto [L(\theta|D_0)]^{a_0}\, \pi_0(\theta|c_0), \tag{4.1}$$

where $\pi_0(\theta|c_0)$ is an *initial prior*. That is, $\pi_0(\theta|c_0)$ is the prior for θ with the historical data being the "current data". The prior parameter c_0 controls the impact of $\pi_0(\theta|c_0)$ on the entire prior, and $0 \leq a_0 \leq 1$ is a scalar prior precision parameter that weights the historical data relative to the likelihood of the current study. Ibrahim and colleagues have called the prior in (4.1) the *power prior* since it is based on exponentiating the likelihood function of the historical data. Small values of a_0 give little prior weight to the historical control data relative to the likelihood of the current study whereas values of a_0 close to 1 give roughly equal weight to the prior and the likelihood of the current study. In particular, the case $a_0 = 1$ corresponds to the update of $\pi_0(\theta|c_0)$ using Bayes theorem; the case $a_0 = 0$ results in no incorporation of historical data and $\pi(\theta|D_0, a_0 = 0) \equiv \pi_0(\theta|c_0)$. We make use of the power prior in this paper to compute the L measure and calibration distribution p_{L_c}.

The prior specification is completed by specifying a prior distribution for a_0. We take a beta prior for a_0, and propose a joint prior distribution for (θ, a_0) of the form

$$\pi(\theta, a_0 \mid D_0) \propto [L(\theta|D_0)]^{a_0}\, \pi_0(\theta|c_0)\; a_0^{\delta_0 - 1}(1 - a_0)^{\lambda_0 - 1}, \tag{4.2}$$

where (δ_0, λ_0) are specified prior hyperparameters. The joint posterior distribution of θ and a_0 is given by

$$p(\theta, a_0|D, D_0) \propto L(\theta|D)\, [L(\theta|D_0)]^{a_0}\, a_0^{\delta_0 - 1}(1 - a_0)^{\lambda_0 - 1}. \tag{4.3}$$

The priors in (4.2) have been proposed and discussed in detail for generalized linear models and survival models; see Ibrahim, Ryan and Chen (1998), Chen, Ibrahim and Yiannoutsos (1999) and Ibrahim and Chen (1998). In these papers, various properties of the power prior $\pi(\theta, a_0 \mid D_0)$ are also derived. We refer the reader to these papers for motivation and detailed discussion of the power prior.

5. Properties of the Calibration Distribution

In this section, we examine some properties of the L measure and the calibration distribution p_{L_c} for the linear model. Consider the linear model

$$y = X\beta + \epsilon \ , \tag{5.1}$$

where β is a $p \times 1$ vector of regression coefficients, X is an $n \times p$ full rank matrix of covariates, and $\epsilon \sim N_n(0, \sigma^2 I)$. For the purposes of our illustrations, and without loss of generality, we take σ^2 to be known. Consider the usual conjugate prior

$$\beta|\sigma^2 \sim N_p(\mu_0, \sigma^2\Sigma_0) \ , \tag{5.2}$$

where (μ_0, Σ_0) are specified hyperparameters. Standard results from Box and Tiao (1973) give the posterior distribution of β as

$$\beta|y \sim N_p(\Lambda\mu_0 + (I_p - \Lambda)\hat{\beta}, \sigma^2(X'X + \Sigma_0^{-1})^{-1}), \tag{5.3}$$

where $\Lambda = (X'X + \Sigma_0^{-1})^{-1}\Sigma_0^{-1}$, I_p denotes the $p \times p$ identity matrix, and $\hat{\beta} = (X'X)^{-1}X'Y$. In addition, the posterior predictive distribution of $z|y$ is given by

$$z|y \sim N_n(X(\Lambda\mu_0 + (I_p - \Lambda)\hat{\beta}), \sigma^2(I_n + X(\Sigma_0^{-1} + X'X)^{-1}X')) \ . \tag{5.4}$$

For the linear model, it can be shown that the L measure has a closed form and is given by

$$L(y, \nu) = n\sigma^2 + \sigma^2 tr(I_p - \Lambda) + \nu(By - X\Lambda\mu_0)'(By - X\Lambda\mu_0), \tag{5.5}$$

where $tr(.)$ denotes the trace operator, $B = I_n - M + X\Lambda(X'X)^{-1}X'$, and $M = X(X'X)^{-1}X'$ is the orthogonal projection operator onto the column space of X, which we denote by $C(X)$. Details of the derivation of (5.5) are given in the Appendix. Moreover, it can be shown that B is a positive definite matrix and $B^{-1} = I_n + X\Sigma_0X'$.

To derive the calibration distribution p_{L_c}, we first need the prior predictive distribution of y. After straightforward algebra,

$$y \sim N_n(X\mu_0, \sigma^2 B^{-1}). \tag{5.6}$$

We are now led to the following theorem which characterizes the exact distribution of the L measure.

Theorem 1. *Under* (5.1)*, the marginal prior predictive distribution of* $L(y, \nu)$ *in* (5.5) *has the same distribution as the random variable* $V = n\sigma^2 + \sigma^2 tr(I_p -$

$\Lambda) + \nu U$, *where* $U = \frac{1}{\sigma^2}\sum_{i=1}^{n} d_{ii}U_i$, *the* U_i *are independent central chi-square random variables each with* 1 *degree of freedom, and the* d_{ii} *are the eigenvalues of the positive definite matrix* B.

The proof of Theorem 1 is given in the Appendix. Now consider a candidate model c and the true model t. Under model j, the prior for β_j is given by

$$\beta_j|\sigma^2 \sim N(\mu_{0j}, \sigma^2\Sigma_{0j}) \ , \tag{5.7}$$

where (μ_{0j}, Σ_{0j}) are specified hyperparameters, $j = c, t$. Under these priors, using the derivation from (5.5), we are led to

$$L_j(y,\nu) = n\sigma^2 + \sigma^2 tr(I_j - \Lambda_j) + \nu(B_j y - X_j\Lambda_j\mu_{0j})'(B_j y - X_j\Lambda_j\mu_{0j})' \ , \tag{5.8}$$

where $\Lambda_j = (X_j'X_j + \Sigma_{0j}^{-1})^{-1}\Sigma_{0j}^{-1}$, $B_j = I_n - M_j + X_j\Lambda_j(X_j'X_j)^{-1}X_j'$, and $M_j = X_j(X_j'X_j)^{-1}X_j'$, $j = c, t$. In addition, using (5.6), the prior predictive distribution of y under the true model t is given by

$$y \sim N_n(X_t\mu_{0t}, \sigma^2 B_t^{-1}) \ , \tag{5.9}$$

where $B_t^{-1} = I_n + X_t\Sigma_{0t}X_t'$.

Theorem 2. *The distribution of* $L_c(y,\nu)$ *with respect to the prior predictive distribution of the true model* t *has the same distribution as the random variable* $V = n\sigma^2 + \sigma^2 tr(I_c - \Lambda_c) + \nu U$, *where* $U = \frac{1}{\sigma^2}\sum_{i=1}^{n} d_{ii}U_i$, *the* U_i *are independent non-central chi-square random variables with* 1 *degree of freedom and non-centrality parameter* λ_i, $i = 1, \ldots, n$. *Here the* d_{ii} *are the eigenvalues of the positive definite matrix* $B_c B_t^{-1} B_c$, *and the* λ_i *are derived explicitly in the Appendix.*

This result is attractive since it gives us an exact closed form expression for the calibration distribution of $L_c(y,\nu)$.

Theorem 3. *The mean* $\mu_c(\nu)$ *of the calibration distribution* p_{L_c} *is given by*

$$\begin{aligned}\mu_c(\nu) = {} & \nu b_{ct}'b_{ct} + \sigma^2(1-\nu)(p_c - p_t) + \sigma^2(1-\nu)(tr(\Lambda_t) - tr(\Lambda_c)) \\ & + \nu\sigma^2 tr(X_c\Lambda_c^2(X_c'X_c)^{-1}X_c'X_t\Sigma_{0t}X_t') + \nu\sigma^2 tr((I_c - M_c)X_t\Sigma_{0t}X_t') \\ & - \nu\sigma^2 tr(\Lambda_c(I_c - \Lambda_c)),\end{aligned} \tag{5.10}$$

where $b_{ct} = (B_c X_t\Lambda_t\mu_{0t} - X_c\Lambda_c\mu_{0c})$.

A proof of Theorem 3 is given in the Appendix. We note that the term b_{ct} in (5.10) can be viewed as a bias component, which equals 0 when $c = t$. We consider several theoretical properties of $\mu_c(\nu)$ for various values of ν and various structures for the models (c, t).

Theorem 4. *Assume* $\Sigma_{0j} = \sigma_0^2 I_j$, $\mu_{0j} = 0$, *and orthogonal covariates, so that* $X_j' X_j = I_j$, $j = c, t$. *Let* $p_j = rank(X_j)$, *and let* $C(X_j)$ *denote the column space of* X_j, $j = c, t$. *Then if*

(i) $C(X_t) \subset C(X_c)$ *and* $\nu < \frac{1+\sigma_0^2}{2+\sigma_0^2}$, *then* $\mu_c(\nu) > 0$;

(ii) $C(X_c) \subset C(X_t)$ *and* $\nu > \frac{1}{2+\sigma_0^2}$, *then* $\mu_c(\nu) > 0$.

Theorem 5. *Assume* $\Sigma_{0j} = \sigma_0^2 I_j$, $\mu_{0j} = 0$, *and orthogonal covariates, so that* $X_j' X_j = I_j$, $j = c, t$. *Let* $p_j = rank(X_j)$, *and let* $C(X_j)$ *denote the column space of* X_j, $j = c, t$. *For any pair of models* (c, t), *if* $\frac{1}{2+\sigma_0^2} < \nu < \frac{1+\sigma_0^2}{2+\sigma_0^2}$, *then* $\mu_c(\nu) > 0$.

Proofs of Theorems 4 and 5 are given in the Appendix. Theorem 5 is a generalization of Theorem 4, and states that for any pair of models, when ν is in the open interval $(\frac{1}{2+\sigma_0^2}, \frac{1+\sigma_0^2}{2+\sigma_0^2})$, the true model t *always* has a smaller L measure value, on average, compared to any other candidate model c. This is a very strong result for linear models since it guarantees that for certain values of ν, the true model always obtains the smaller L measure value regardless of its relation to the candidate model. Theorems 4 and 5 remain valid when $\mu_{0j} \neq 0$ and Σ_{0j} is allowed to be more general. From Theorem 5, when $\sigma_0^2 \to \infty$, the interval converges to $(0, 1)$. Then, under noninformative priors, the true model always has the smallest L measure value on average for all possible values of ν.

An interesting special case has $\nu = 1$, which corresponds to the criterion of Ibrahim and Laud (1994) and Laud and Ibrahim (1995). Under this case, interesting properties of the L measure are obtained.

Theorem 6. *Suppose* $\nu = 1$ *in* (5.5), X_j *is an arbitrary covariate matrix and* Σ_{0j} *is diagonal,* $j = c, t$. *If*

(i) $C(X_c) \subset C(X_t)$ *then* $\mu_c(1) > 0$;

(ii) $C(X_t) \subset C(X_c)$ *and* $\mu_{0j} = 0$, $j = c, t$, *then* $\mu_c(1) < 0$;

(iii) $C(X_t) \subset C(X_c)$ *and* $\mu_{0j} \neq 0$, *for either* $j = c, t$, *then* $\mu_c(1)$ *may be positive or negative.*

A proof of Theorem 6 is given in the Appendix. Parts (i) and (ii) state that, when $\nu = 1$ and $\mu_{0j} = 0$, $j = c, t$, then the larger model always has smaller L measure value on average, regardless of whether it is the true model or not. This result is valid for any covariate matrices X_c and X_t. Part (iii) states that when the true model is contained in the candidate model, then the sign of $\mu_c(1)$ is determined by the prior hyperparameters, specifically by μ_{0j}, $j = c, t$. Thus in this case, for certain choices of μ_{0j}, $\mu_c(1)$ can either be positive or negative, implying that the L measure for the candidate model can be smaller or larger than the L measure for the true model. The results in (i), (ii), and (iii) of Theorem

6 imply that $\nu = 1$ may not be the most suitable choice for the L measure for situations involving nested models.

6. Examples

Example 1. Simulation Study

We consider a simulation study to compare several models using the proposed L measure and calibration distribution for parametric survival models with right censoring. We generate the current data y_i from the Weibull model

$$p(y_i|\alpha, \lambda_i) = \alpha^{y_i - 1} \exp\{\lambda_i - y_i^{\alpha} \exp(\lambda_i)\}, \tag{6.1}$$

where $\lambda_i = \beta_0 + \beta_1 x_i$, $i = 1, \ldots, n$, $n = 200$, $\alpha = 2$, $\beta = (\beta_0, \beta_1) = (1, -1)$, and 10% of the observations are randomly right censored with censoring times $t_i = 0.75 * y_i$. The covariates x_i are *i.i.d.* $N(0, 1)$ variates. We generate historical data $D_0 = (n_0, y_0, X_0)$ using the Weibull model with $n_0 = 100$, $\alpha = 2$, $\beta = (0.9, -0.9)$, $x_{0i} \sim N(0, 1)$, and 10% censoring with $t_{i0} = 0.75 * y_{0i}$. Also, we take the initial prior to be $\pi_0(\beta, \alpha|c_0) = \pi_0(\beta|c_0)\pi_0(\alpha)$, where $\pi_0(\beta|c_0)$ is $N_2(0, c_0 I)$, $c_0 = 4$, and $\pi_0(\alpha)$ is a gamma density with shape and scale parameters $(1, 0.5)$, respectively. Here, $N_p(\mu, \Sigma)$ denotes the p dimensional multivariate normal distribution with mean μ and covariance matrix Σ.

We consider three models: 1) the Weibull model from which the data were generated, 2) the Weibull model with a random α parameter, and 3) the exponential model ($\alpha = 1$). We specify a vague gamma prior for α in model 2). All L measure and p_{L_c} computations are based on survival times transformed to logarithms. Table 1 below gives posterior summaries for the three models using a $beta(50, 50)$ prior for a_0.

Table 1. Posterior summaries for simulated data.

	True model	Weibull (random α)	exponential
$E(a_0\|D, D_0)$	0.46	0.45	0.39
$sd(a_0\|D, D_0)$	0.05	0.05	0.05
$E(\beta_0\|D, D_0)$	0.964	0.929	0.559
$sd(\beta_0\|D, D_0)$	0.067	0.075	0.068
95% HPD for β_0	(0.831, 1.096)	(0.783, 1.075)	(0.424, 0.691)
$E(\beta_1\|D, D_0)$	-1.093	-1.036	-0.549
$sd(\beta_1\|D, D_0)$	0.077	0.092	0.075
95% HPD for β_1	$(-1.246, -0.943)$	$(-1.213, -0.853)$	$(-0.695, -0.403)$
$E(\alpha\|D, D_0)$	–	1.894	–
$sd(\alpha\|D, D_0)$	–	0.094	–
95% HPD for α	–	(1.710, 2.083)	–

Table 2. L measure and calibration summaries for simulated data.

model	*L measure*	$\mu_c(1/2)$	50% HPD	95% HPD
1*	131.28	-	-	-
2	145.81	14.50	(-6.90, 14.21)	(-18.15, 68.92)
3	394.58	262.37	(264.82, 268.96)	(237.56, 274.13)

* True model

Table 2 shows results of the L measure and the calibration distributions p_{L_c} for Models 2 and 3 using $\nu = 1/2$. The first row of Table 2 represents the true model in which the L measure value is given. The remaining entries in the first row of Table 2 are blank due to the definition of the calibration distribution. Figure 1 shows a superimposed plot of the calibration distributions for Models 2 and 3. The horizontal axis in Figure 1 are values of $\hat{D}(y_{obs}, 1/2)$ and the vertical axis are values of the density $\hat{p}_{L_c}$. From Table 1, we see that the posterior estimates for Models 1 and 2 are close, but are quite different from those of Model 3. From Table 2, we see that the criterion minimizing model is in fact the true model with an observed L measure value of 131.3. The observed L measure values for the other models are larger than for the true model, and for the exponential model it is substantially larger. A formal comparison of the observed L measure values for the candidate models and the true model can be assessed by an examination of p_{L_c}. From Table 2, we see that Model 2 is much closer to the true model than Model 3, as $\mu_2(1/2) = 14.50$ and $\mu_3(1/2) = 262.37$. From Table 2, we observe that the HPD intervals for p_{L_2} contain 0 whereas the HPD intervals for p_{L_3} do not. The HPD intervals for p_{L_3} are much narrower than those for p_{L_2}, since p_{L_3} has smaller dispersion than p_{L_2}. By taking α random and specifying a prior distribution for it, we introduce more variability in the L measure, and hence the resulting calibration distribution. This results in wider HPD intervals. We note that all HPD intervals presented in Tables 1 and 2 as well as the subsequent tables were computed using a Monte Carlo method developed by Chen and Shao (1999). Figure 1 shows the calibration distributions for Models 2 and 3. Figure 1 presents a useful summary of the properties of the L measure for Models 2 and 3 compared to the true model, and clearly shows that Model 3 is quite different from the true model, and Models 2 and 3 are quite different from each other.

An analysis of these data was also conducted with a $beta(10, 10)$ prior for a_0, as well as with a point mass prior at $a_0 = 0$. Similar patterns to those of Table 2 were observed for the L measure and the calibration distribution p_{L_c}. For example, for $a_0 \sim beta(10, 10)$, the observed L measure values for Models i), ii), and iii) were 131.39, 142.84, and 396.05, respectively, and p_{L_c} was nearly identical to that of Figure 1. A similar phenomenon was observed for the case $a_0 = 0$ with probability 1.

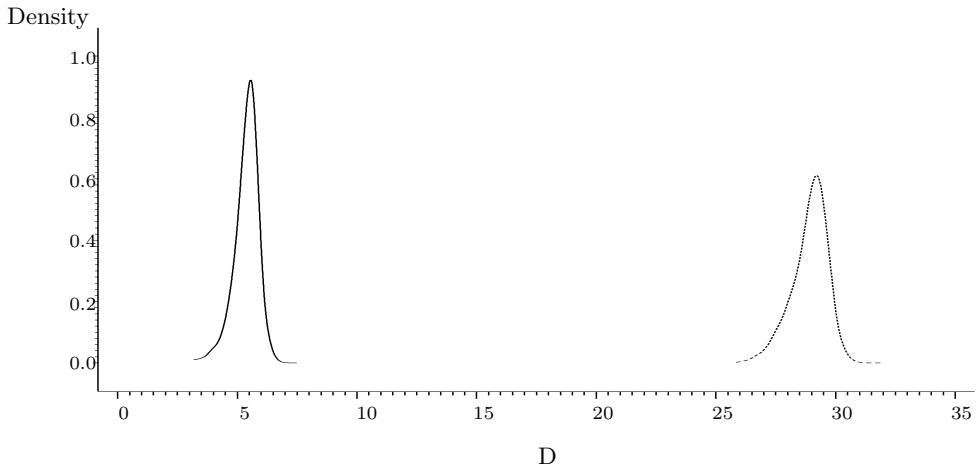

Figure 1. Calibration distributions for simulated data: the solid curve is Model 2), the dashed curve is Model 3).

Example 2. AIDS Data

We revisit the AIDS study ACTG036 discussed in Section 1. We use the data from a similar AIDS study, ACTG019, as historical data. Chen, Ibrahim and Yiannoutsos (1999) discuss these data in detail, and demonstrate prior elicitation with historical data and Bayesian variable selection in logistic regression.

The ACTG019 study was a double blind placebo-controlled clinical trial comparing zidovudine (AZT) to placebo in persons with CD4 counts less than 500. The results of this study were published in Volberding et al. (1990). The sample size for this study, excluding cases with missing data, was $n_0 = 823$. The response variable (y_0) for these data is binary with 1 indicating death, development of AIDS, or AIDS related complex (ARC), and 0 indicating other. Several covariates were also measured. The ones we use here are CD4 count (x_{01}) (cell count per mm^3 of serum), age (x_{02}), treatment, (x_{03}), and race (x_{04}). The covariates CD4 count and age are continuous, whereas the other covariates are binary. The same covariates as those of ACTG036 were measured for ACTG019, and thus the full model consists of these 4 covariates. We consider a logistic regression for the response y as a function of the covariates (see Chen, Ibrahim and Yiannoutsos (1999)).

We compute the L measure and calibration distribution for three competing logistic regression models:

Model 1): (x_1, x_2, x_3),

Model 2): (x_1, x_2, x_3, x_4) (the full model)

Model 3): (x_1).

For each model, we use a joint prior for (β, a_0) as in (4.2), where $a_0 \sim beta(20, 20)$, and $\pi_0(\beta|c_0)$ is a multivariate normal density with mean 0 and

covariance matrix c_0W_0. We use $c_0 = 100$ and take W_0 to be a diagonal matrix consisting of the asymptotic standard errors of the maximum likelihood estimates of β based on the historical data D_0. We have suppressed the dependence of each prior on the model for ease of exposition. We note that $\pi(\beta, a_0|D_0)$ and $\pi_0(\beta|c_0)$ change dimension for each model under consideration, and we have suppressed this label here as well. Table 3 shows summaries of the L measure and p_{L_c} for the models using $\nu = 1/2$. The posterior mean and standard deviation of a_0 for each model is also given in Table 3. Figure 2 shows a superimposed plot of the calibration distribution p_{L_c} for the candidate models. The horizontal axis in Figure 2 are values of $\hat{D}(y_{obs}, 1/2)$ and the vertical axis are values of $\hat{p}_{L_c}$.

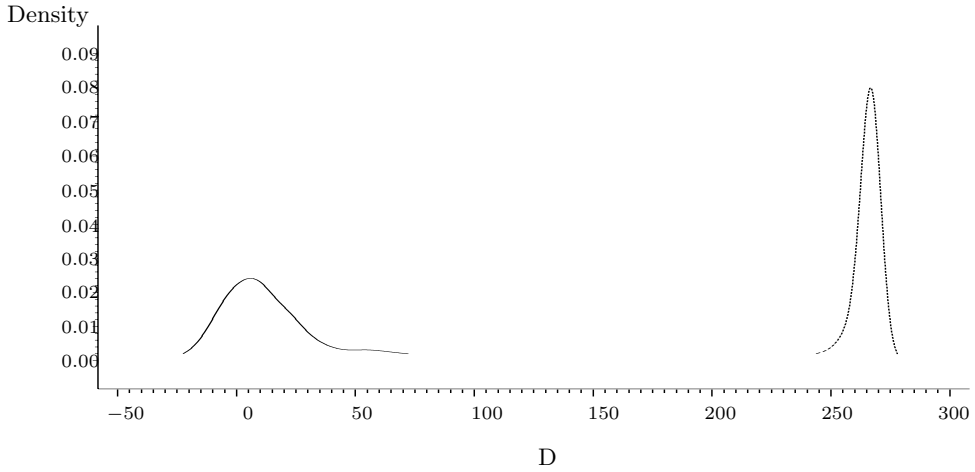

Figure 2. Calibration distributions for AIDS data: the solid curve is Model 1), the dashed curve is Model 3).

Table 3. L measure and calibration summaries for AIDS data.

model	*L measure*	$\mu_c(1/2)$	50% HPD	95% HPD	$E(a_0\|D, D_0)$	$sd(a_0\|D.D_0)$
1	16.37	0.151	(0.045, 0.217)	(-0.133, 0.451)	0.42	0.05
2*	16.30	-	-	-	0.43	0.05
3	16.82	1.729	(0.782, 1.863)	(0.246, 3.673)	0.41	0.05

* Criterion minimizing model

From Table 3, we see that the criterion minimizing model is Model 2), which is the full model, with an observed L measure value of 16.30. We also see that all models are nearly identical in their criterion values. Based on the criterion value alone, it is impossible to distinguish between Models 1 and 3 in relation to the criterion minimizing Model 2. Thus, we must turn to the calibration distribution p_{L_c} for a better assessment of the two models. From Table 3, we see that $\mu_1(1/2) = 0.151$, $\mu_3(1/2) = 1.729$, and the 95% HPD intervals for Model

1 contain 0 whereas the HPD intervals for Model 3 do not. This implies that, although the L measure values for Models 1 and 3 are close, Model 1 is much closer to the criterion minimizing model than Model 3. In addition, Model 1 has $\mu_1(1/2)$ close to 0, and thus perhaps could serve as a suitable parsimonious model for these data. In Figure 2, we see that the calibration distribution for Model 1 is nearly symmetric about 0.15, and has much smaller dispersion than the calibration distribution for Model 3. Based on parsimony, it appears that the (x_1, x_2, x_3) model is a suitable one.

Example 3. Breast Cancer Data

We consider the breast cancer data from Finkelstein and Wolf (1985), which consists of a data set of (case-2) interval censored data. In this data set, 46 early breast cancer patients receiving only radiotherapy (covariate value $x = 0$) and 48 patients receiving radio-chemotherapy ($x = 1$) were monitored for cosmetic change through weekly clinic visits. Some weekly visits were missed. The data on survival time are typically recorded as, for example, [7,18] (at the 7th week clinic-visit patient had shown no change, and in the next clinic visit at the 18th week the patient's tissue showed that the change had already occurred). We are interested in the effect of the covariate x on the survival time y. Sinha et al. (1999) consider a semi-parametric Bayesian analysis of these data using three models based on a discretized version of the Cox model (Cox (1972)). Specifically, the hazard, $\lambda(y|x)$, is taken to be a piecewise constant function with $\lambda(y|x) = \lambda_j \theta_j^x$ for $y \in I_j$, where $\theta_j = e^{\beta_j}$, $I_k = (a_{j-1}, a_j]$ for $j = 1, 2, \ldots, g$, $0 = a_0 < a_1 < \cdots < a_g = \infty$, and g is the total number of grid intervals. The grid can be taken sufficiently fine to approximate any hazard function for all practical purposes. The three models are defined by

Model 1): (i) $\lambda_j \overset{i.i.d.}{\sim} Gamma(\eta_j, \gamma_j)$ for $j = 1, \ldots, g$; (ii) $\beta_j \equiv \beta$, $j = 1, \ldots, g$, and $\beta \sim N(\beta_0, w_0^2)$;

Model 2): (i) λ_j's has same prior as in model 1); (ii) $\beta_{j+1} \mid \beta_1, \ldots, \beta_{g-1} \sim N(\beta_j, w_j^2)$ for $j = 0, \ldots, g-1$;

Model 3): (i) $\alpha_{j+1} \mid \alpha_1, \ldots, \alpha_j \sim N(\alpha_j, v_j^2)$, for $j = 0, 1, \ldots, g-1$; (ii) Same as in Model 2), $\alpha_j = \log(\lambda_j)$.

Our purpose here is to compute the L measure and the calibration distribution for the three models, using noninformative priors. We use the same values of the prior parameters as Sinha et al. (1999) as follows.

	η_k	γ_k	β_0	w_k	w_0	v_k	α_0	v_0
Model 1)	0.2	0.4	0	–	2.0	–	–	–
Model 2)	0.2	0.4	0	1.0	2.0	–	–	–
Model 3)	–	–	0	–	2.0	1.0	-0.1	2.0

Table 4 shows the results using $\nu = 1/2$. Model 1) is the criterion minimizing model with L measure value of 80.45. Since $\mu_2(1/2) = 5.36$ and $\mu_3(1/2) = 28.91$, Model 2 is much closer to the criterion minimizing model than Model 3. From Figure 3, we see that there is a wide separation between p_{L_2} and p_{L_3}, and that p_{L_2} has smaller dispersion than p_{L_3}. The HPD intervals for Models 2 and 3 do not contain 0. We conclude that both Models 2 and 3 are different from one another and different from the criterion minimizing model. We note that other choices of prior parameters yielded similar L measure values and calibration distributions.

Table 4. L measure and calibration summaries for breast cancer data.

model	*L measure*	$\mu_c(1/2)$	50% HPD	95% HPD
1*	80.45	-	-	-
2	87.24	5.36	(5.23, 5.83)	(4.24, 6.34)
3	113.54	28.91	(28.71, 29.61)	(27.23, 30.23)

* Criterion minimizing model

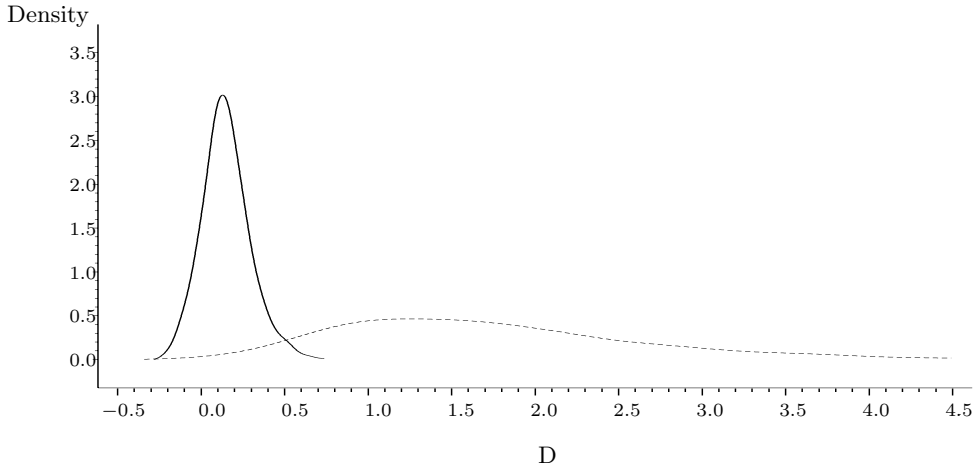

Figure 3. Calibration distributions for breast cancer data: The solid curve is Model 2), the dashed curve is Model 3).

Acknowledgements

The authors wish to thank the Editor, the Associate Editor, and a referee for their helpful comments and suggestions, which led to an improvement in this article. This work was partially supported by NSF grant No. DMS-9702172, and NIH grants #CA 70101, #CA 74015, and R29-CA69222.

Apendix. Proofs of Theorems

By the definition of $L(y, \nu)$ in (2.7), and using the posterior predictive dis-

tribuion $p(z|y)$ in (5.4), it can be shown that

$$L(y,\nu) = n\sigma^2 + tr(I-\Lambda) + \nu(By - X\Lambda\mu_0)'(By - X\Lambda\mu_0),$$

where B and M are defined below (5.5).

Proof of Theorem 1. Let $\tilde{y} = By - X\Lambda\mu_0$. We first need to obtain the prior predictive distribution of $\tilde{y}$. Using (5.6), we have $E(By - X\Lambda\mu_0) = 0$ and $\mathrm{Cov}(By - X\Lambda\mu_0) = \sigma^2 B$. Thus the prior predictive distribution of $\tilde{y}$ is $\tilde{y} \sim N_n(0, \sigma^2 B)$. Now we can write

$$\tilde{y}'\tilde{y} = \tilde{y}'\left(B^{-1/2}B^{1/2}B^{1/2}B^{-1/2}\right)\tilde{y} = Z'BZ = Z'(PDP')Z = W'DW \ , \quad \text{(A.1)}$$

where $Z = (B^{-1/2}\tilde{y}) \sim N_n(0, \sigma^2 I_n)$ and $B = PDP'$ is the spectral decomposition of B. Thus, since $W \sim N_n(0, \sigma^2 I)$, letting W_i denote the ith component of the vector W, we can write (A.1) as $W'DW = \sum_{i=1}^n d_{ii}W_i^2 \sim \frac{1}{\sigma^2}\sum_{i=1}^n d_{ii}U_i$, where the U_i are *i.i.d.* central chi-square random variables each with one degree of freedom. This completes the proof.

Proof of Theorem 2. We follow a proof similar to that of Theorem 1. Let $\tilde{Y} = B_c y - X_c\Lambda_c\mu_{0c}$. Then under the true model t, using (5.9), we have $E_t(\tilde{y}) = B_c X_t\mu_{0t} - X_c\Lambda_c\mu_{0c}$ and $\mathrm{Cov}_t(\tilde{y}) = B_c B_t^{-1} B_c$, so $\tilde{y} \sim N_n(B_c X_t\mu_{0t} - X_c\Lambda_c\mu_{0c}, \sigma^2 B_c B_t^{-1} B_c) = N_n(\mu_{ct}, \sigma^2\Sigma_{ct})$, say. We can write

$$\tilde{y}'\tilde{y} = \tilde{y}'\left(\Sigma_{ct}^{-1/2}\Sigma_{ct}^{1/2}\Sigma_{ct}^{1/2}\Sigma_{ct}^{-1/2}\right)\tilde{y} = Z'\Sigma_{ct}Z = Z'(PDP')Z = W'DW \ , \quad \text{(A.2)}$$

where $Z = \Sigma_{ct}^{-1/2}\tilde{y}$, $W = P'Z$, P is the orthogonal matrix consisting of the eigenvectors of Σ_{ct}, and $D = (d_{ii})$ is the diagonal matrix of eigenvalues of Σ_{ct}, denoted by d_{ii}. Since $W \sim N(P'\Sigma_{ct}^{-1/2}\mu_{ct}, \sigma^2 I)$, it follows from (A.2) that $W'DW = \sum_{i=1}^n d_{ii}W_i^2 \sim \frac{1}{\sigma^2}\sum_{i=1}^n d_{ii}U_i$, the U_i are independent non-central chi-square random variables each with one degree of freedom and non-centrality parameter $\lambda_i = \frac{\mu_i^2}{2\sigma^2}$, μ_i the ith component of the vector $P'\Sigma_{ct}^{-1/2}\mu_{ct}$.

Proof of Theorem 3. Using (5.8) along with (5.9), we have

$$\begin{aligned}\mu_c(\nu) = E_t(L_c(y,\nu) - L_t(y,\nu)) &= n\sigma^2 - n\sigma^2 + \sigma^2\left(tr(I_c - \Lambda_c) - tr(I_t - \Lambda_t)\right) \\ &+\nu E_t\left[(B_c y - X_c\Lambda_c\mu_{0c})'(B_c y - X_c\Lambda_c\mu_{0c})\right] \\ &-\nu E_t\left[(B_t y - X_t\Lambda_t\mu_{0t})'(B_t y - X_t\Lambda_t\mu_{0t})\right].\end{aligned}$$

Following along the same lines as before, under the true model, we have $E_t(B_t y - X_t\Lambda_t\mu_{0t}) = 0$, $Cov_t(B_t Y - X_t\Lambda_t\mu_{0t}) = B_t$, so $B_t y - X_t\Lambda_t\mu_{0t} \sim N_n(0, \sigma^2 B_t)$. Similarly, under (5.9), we have $B_c y - X_c\Lambda_c\mu_{0c} \sim N_n(B_c X_t\Lambda_t\mu_{0t} - X_c\Lambda_c\mu_{0c}, \sigma^2 (B_c B_t^{-1} B_c))$. We find

$$E_t\left[(B_c y - X_c\Lambda_c\mu_{0c})'(B_c y - X_c\Lambda_c\mu_{0c})\right] = \sigma^2 tr(B_c B_t^{-1} B_c) + b_{ct}'b_{ct},$$

where $b_{ct} = B_c X_t \Lambda_t \mu_{0t} - X_c \Lambda_c \mu_{0c}$. Since $tr(B_t) = n - p_t + tr(\Lambda_t)$ and $tr(B_c B_t^{-1} B_c)$ $= n - p_c + tr(\Lambda_c^2) + tr((I_n - M_c) X_t \Sigma_{0t} X_t') + tr(X_c \Lambda_c^2 (X_c' X_c)^{-1} X_c' X_t \Sigma_{0t} X_t')$, we get

$$\begin{aligned} & E_t(L_c(y,\nu) - L_t(y,\nu)) \\ &= \sigma^2(1-\nu)(p_c - p_t) + \sigma^2(1-\nu) tr(\Lambda_t) - \sigma^2(1-\nu) tr(\Lambda_c) - \sigma^2 \nu tr(\Lambda_c(I_c - \Lambda_c)) \\ &\quad + \sigma^2 \nu tr((I_c - M_c) X_t \Sigma_{0t} X_t') + \sigma^2 \nu tr(X_c \Lambda_c^2 (X_c' X_c)^{-1} X_c' X_t \Sigma_{0t} X_t') + \nu b_{ct}' b_{ct}. \quad \text{(A.3)} \end{aligned}$$

This completes the proof.

Proof of Theorem 4. (i) We first consider $C(X_t) \subset C(X_c)$ and write $X_c = (X_c *, X_t)$, where $\text{rank}(X_c *) = p_c - p_t$. Since $X_c = (X_c *, X_t)$, we can write $X_t \Sigma_{0t} X_t' = X_c \Sigma_{0c} X_c' - X_{c^*} \Sigma_{0c^*} X_{c^*}'$. To obtain $\mu_c(\nu)$ defined in (5.10), we first note that if $\mu_{0j} = 0$, for $j = c, t$, then $b_{ct} = 0$. Using this decomposition, we have $tr(\Lambda_c) = \frac{p_c}{1+\sigma_0^2}$, $tr(\Lambda_t) = \frac{p_t}{1+\sigma_0^2}$, and $tr(\Lambda_c(I_c - \Lambda_c)) = \frac{p_c \sigma_0^2}{(1+\sigma_0^2)^2}$. Since $C(X_t) \subset C(X_c)$, $(I_c - M_c) X_t \Sigma_{0t} X_t' = 0$, and $tr(X_c \Lambda_c^2 (X_c' X_c)^{-1} X_c' X_t \Sigma_{0t} X_t') = \frac{\sigma_0^2 p_t}{(1+\sigma_0^2)^2}$. Substituting these expressions into the formula for $\mu_c(\nu)$, we get $\mu_c(\nu) = \sigma^2(1-\nu)(p_c - p_t) - \sigma^2(1-\nu)\frac{(p_c - p_t)}{(1+\sigma_0^2)} - \sigma^2 \nu \frac{\sigma_0^2 (p_c - p_t)}{(1+\sigma_0^2)^2}$. Now $\mu_c(\nu) > 0$ if and only if $\nu < \frac{1+\sigma_0^2}{2+\sigma_0^2}$.

(ii) When $C(X_c) \subset C(X_t)$, $p_t > p_c$. We partition $X_t = (X_c, X_{t^*})$, where $\text{rank}(X_{t^*}) = p_t - p_c$. In this case, we can write $X_t \Sigma_{0t} X_t' = X_c \Sigma_{0c} X_c' + X_{t^*} \Sigma_{0t^*} X_{t^*}'$. In addition, since X_c and X_{t^*} are orthogonal, we have $X_c' X_{t^*} = 0$. Thus, we have $tr(X_c \Lambda_c^2 (X_c' X_c)^{-1} X_c' X_t \Sigma_{0t} X_t') = \frac{p_c \sigma_0^2}{(1+\sigma_0^2)^2}$ and $tr((I_n - M_c) X_t \Sigma_{0t} X_t') = \sigma_0^2 (p_t - p_c)$. Thus, $\mu_c(\nu) = -\sigma^2(1-\nu)(p_t - p_c) + \frac{(1-\nu)}{1+\sigma_0^2}(p_t - p_c) + \sigma^2 \nu \sigma_0^2 (p_t - p_c)$, and $\mu_c(\nu) > 0$ if and only if $\nu > \frac{1}{2+\sigma_0^2}$.

Proof of Theorem 5. Let $p^* = \dim(C(X_c) \cap C(X_t))$, and write $X_c = (X_{p^*}, X_{c^*})$, $X_t = (X_{p^*}, X_{t^*})$, $p_{c^*} = \text{rank}(X_{c^*})$, $p_{t^*} = \text{rank}(X_{t^*})$, and $p^* = \text{rank}(X_{p^*})$. Thus $p_{c^*} = p_c - p^*$ and $p_{t^*} = p_t - p^*$. Now we can write $X_t \Sigma_{0t} X_t' = \sigma_0^2 (X_c X_c' - X_{c^*} X_{c^*}')$. After some algebra it can be shown that $\mu_c(\nu) = \frac{\sigma^2 \sigma_0^2}{1+\sigma_0^2} \left((2+\sigma_0^2)\nu - 1\right)(p_t - p^*) + \frac{\sigma^2 \sigma_0^2}{1+\sigma_0^2}(p_c - p^*)\left(1 - \nu\left(\frac{2+\sigma_0^2}{1+\sigma_0^2}\right)\right)$, and so $\mu_c(\nu) > 0$ if and only if $\frac{1}{2+\sigma_0^2} < \nu < \frac{1+\sigma_0^2}{2+\sigma_0^2}$.

Proof of Theorem 6. (i) Set $\nu = 1$ in (A.3) and consider $C(X_c) \subset C(X_t)$. We can write $X_t = (X_c, X_t^*)$ and $\Sigma_{0t} = \begin{pmatrix} \Sigma_{0c} & 0 \\ 0 & \Sigma_{0t^*} \end{pmatrix}$. Thus $X_t \Sigma_{0t} X_t' = X_c \Sigma_{0c} X_c' + X_{t^*} \Sigma_{0t^*} X_{t^*}'$. Also, we note that $I_c - \Lambda_c = (X_c' X_c + \Sigma_{0c}^{-1})^{-1} X_c' X_c$. Therefore, $\mu_c(1) = \sigma^2 tr((I_c - M_c) X_{t^*} \Sigma_{0t} X_{t^*}') + \sigma^2 tr(X_c \Lambda_c^2 (X_c' X_c)^{-1} X_c' X_{t^*} \Sigma_{0t^*} X_{t^*}') + b_{ct}' b_{ct} > 0$. The first two terms in $\mu_c(1)$ are positive since they are both of the form $tr(AB)$

where A and B are positive definite matrices. Clearly, $b'_{ct}b_{ct} > 0$ for $b_{ct} \neq 0$. This proves part (i).

(ii) If $C(X_t) \subset C(X_c)$, write $X_c = (X_c^*, X_t)$ and $\Sigma_{0c} = \begin{pmatrix} \Sigma_{0c^*} & 0 \\ 0 & \Sigma_{0t} \end{pmatrix}$. Then $(I_c - M_c)X_t = 0$, and $X_t\Sigma_{0t}X_t' = X_c\Sigma_{0c}X_c' - X_{c^*}\Sigma_{0c^*}X_{c^*}'$. Thus we have

$$\mu_c(1) = -\sigma^2 tr(X_c\Lambda_c^2(X_c'X_c)^{-1}X_c'X_{c^*}\Sigma_{0c^*}X_{c^*}') + b'_{ct}b_{ct} \ . \tag{A.4}$$

The term $tr(X_c\Lambda_c^2(X_c'X_c)^{-1}X_c'X_{t^*}\Sigma_{0c^*}X_{c^*}')$ in (A.7) is always positive since it is of the form $tr(AB)$ where A and B are positive definite matrices. We make use of the fact that $tr(X_c\Lambda_c^2(X_c'X_c)^{-1}X_c'(X_c\Sigma_{0c}X_c')) = tr(X_c\Lambda_c(X_c'X_c + \Sigma_{0c}^{-1})^{-1}X_c')$. Thus (A.7) can take on positive or negative values depending on the value of $b'_{ct}b_{ct}$. Specifically, if $\mu_{0c} = 0$ and $\mu_{0t} = 0$, $\mu_c(1) < 0$ in (A.7). However, if either $\mu_{0c} \neq 0$ or $\mu_{0t} \neq 0$, then $\mu_c(1) > 0$ is possible. This completes the proof.

References

Bayarri, M. J. and Berger, J. O. (1999). Quantifying surprise in the data and model verification. In *Bayesian Statistics* **6** (Edited by J. M Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 53-82. Oxford University Press, Oxford.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley, New York.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* **143**, 383-430.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, MA.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London.

Chen, M.-H., Ibrahim, J. G. and Yiannoutsos, C. (1999). Prior elicitation, variable selection and Bayesian computation for logistic regression models. *J. Roy. Statist. Soc. Ser. B* **61**, 223-242.

Chen, M.-H. and Shao, Q.-M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *J. Comput. Graph. Statist.* **8**, 69-92.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In *Bayesian Statistics* **6** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 157-185. Oxford University Press, Oxford.

Cox, D. R. (1972). Regression models and life tables. *J. Roy. Statist. Soc. Ser. B* **34**, 187-220.

Finkelstein, D. M. and Wolf, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933-945.

Geisser, S. (1993). *Predictive Inference: An Introduction.* Chapman and Hall, London.

Gelfand, A. and Ghosh, S. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85**, 1-13.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995). *Bayesian Data Analysis.* Chapman and Hall, London.

Gelman, A. Meng, X. L. and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sinica* **6**, 733-807.

George, I. E. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881-889.

George, I. E., McCulloch, R. E. and Tsay, R. S. (1995). Two approaches to Bayesian model selections with applications. In *Bayesian Analysis in Econometrics and Statistics - Essays in Honor of Arnold Zellner* (Edited by D. A. Berry, K. A. Chaloner and J. K. Geweke), 339-348. John Wiley, New York.

Ibrahim, J. G. and Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazards models. *Sankhya Ser. B* **60**, 48-64.

Ibrahim, J. G. and Laud, P. W. (1994). A predictive approach to the analysis of designed experiments. *J. Amer. Statist. Assoc.* **89**, 309-319.

Ibrahim, J. G., Ryan, L.-M. and Chen, M.-H. (1998). Use of historical controls to adjust for covariates in trend tests for binary data. *J. Amer. Statist. Assoc.* **93**, 1282-1293.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773-795.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive model selection. *J. Roy. Statist. Soc. Ser. B* **57**, 247-262.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89**, 1535-1546.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models.* 2nd edition. Chapman and Hall, London.

Merigan, T. C., Amato, D. A., Balsley, J., Power, M., Price, W. A., Benoit, S., Perez-Michael, A., Brownstein, A., Kramer, A. S., Brettler, D., Aledort, L., Ragni, M. V., Andes, A. W., Gill, J. C., Goldsmith, J., Stabler, S., Sanders, N., Gjerset, G., Lusher, J. and the NHF-ACTG036 Study Group (1991). Placebo-controlled trial to evaluate zidovudine in treatment of human immunodeficiency virus infection in asymptomatic patients with hemophilia. *Blood* **78**, 900-906.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.* **92**, 179-191.

Raftery, A. E., Madigan, D. and Volinsky, C. T. (1995). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics* **5** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 323-350. Oxford University Press, Oxford.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Sinha, D. and Dey, D. K. (1997). Semiparametric Bayesian analysis of survival data. *J. Amer. Statist. Assoc.* **92**, 1195-1212.

Sinha, D., Chen, M.-H. and Ghosh, S. K. (1999). Bayesian analysis and model selection for interval-censored survival data. *Biometrics* **55**, 585-590.

Spiegelhalter, D. J., Best, N. G. and Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Submitted.

Volberding, P. A., Lagakos, S. W., Koch, M. A., Pettinelli, C., Myers, M. W., Booth, D. K., Balfour, H. H., Reichman, R. C., Bartlett, J. A., Hirsch, M. S., Murphy, R. L., Hardy, D., Soeiro, R., Fischl, M. A., Bartlett, J. G., Merigan, T. C., Hylsop, N. E., Richman, D. D., Valentine, F. T., Corey, L. and the AIDS Clinical Trials Group of the National Institute of Allergy and Infectious Diseases (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. *New England Journal of Medicine* **322**, 941-949.

Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute.

Department of Mathematical Sciences, Worcester Polytechnic Institute.

Department of Biometry and Epidemiology, Medical University of South Carolina.

(Received January 2000; accepted August 2000)