BIOMETRICS 56, 256–262
March 2000

Bayesian Information Criterion for Censored Survival Models

Chris T. Volinsky

AT&T Labs, 180 Park Avenue, Building 103, Room D-259,
Florham Park, New Jersey 07932, U.S.A.
email: volinsky@research.att.com

and

Adrian E. Raftery

University of Washington, Seattle, Washington, U.S.A.

SUMMARY. We investigate the Bayesian Information Criterion (BIC) for variable selection in models for censored survival data. Kass and Wasserman (1995, *Journal of the American Statistical Association* **90,** 928–934) showed that BIC provides a close approximation to the Bayes factor when a unit-information prior on the parameter space is used. We propose a revision of the penalty term in BIC so that it is defined in terms of the number of uncensored events instead of the number of observations. For a simple censored data model, this revision results in a better approximation to the exact Bayes factor based on a conjugate unit-information prior. In the Cox proportional hazards regression model, we propose defining BIC in terms of the maximized partial likelihood. Using the number of deaths rather than the number of individuals in the BIC penalty term corresponds to a more realistic prior on the parameter space and is shown to improve predictive performance for assessing stroke risk in the Cardiovascular Health Study.

KEY WORDS: Bayes factor; Cox proportional hazards model; Exponential distribution; Partial likelihood; Variable selection.

1. Introduction

Many epidemiological studies assess risk factors for death or for the onset of a disease. A set of candidate risk factors is identified, and the data are used to winnow this down to a subset to be used for risk assessment and prediction. One common approach to this model selection task is a Bayesian one, based on the posterior probabilities of the models defined by the possible subsets of the risk factors. The posterior model probabilities can be approximated using BIC, a simple expression that involves only the maximized likelihood, the sample size, and the number of risk factors in the model. We show here that, when there is censoring, an improved approximation is obtained by replacing the total sample size by the number of uncensored cases. This revised approximation preserves, and may improve, the asymptotic properties of the BIC approximation and also provides better results in practice for some standard epidemiological data sets. Our research was motivated by the Cardiovascular Health Study, a study of risk factors for cardiovascular disease in older adults.

The Bayesian framework for hypothesis testing uses Bayes factors to quantify the evidence for one hypothesized model against another (Kass and Raftery, 1995). Schwarz (1978) derived the Bayesian Information Criterion (BIC) as a large-sample approximation to twice the logarithm of the Bayes factor. For a model M_j parameterized by an m_j-dimensional vector θ_j,

$$\text{BIC} = -2\left\{\ell_j(\hat{\theta}_j) - \ell_0(\hat{\theta}_0)\right\} + (m_j - m_0)\log(n), \quad (1)$$

where $\ell_j(\hat{\theta}_j)$ and $\ell_0(\hat{\theta}_0)$ are the maximized likelihoods under M_j and a reference model M_0, whose parameter has dimension m_0, where n is the sample size. With nested models, BIC equals the standard likelihood ratio test statistic minus a complexity penalty that depends on the degrees of freedom of the test of M_0 against M_j. BIC provides an approximation to the Bayes factor, which can readily be computed from the output of standard statistical software packages (Kass and Raftery, 1995; Raftery, 1995).

The derivation of BIC involves a Laplace approximation to the Bayes factor and ignores terms of constant order, so the difference between BIC and twice the log Bayes factor does not vanish asymptotically in general, although it becomes inconsequential in large samples. However, Kass and Wasserman (1995) show that, under certain nonrestrictive regularity conditions, the difference between BIC and twice the log Bayes factor does tend to zero for a specific choice of prior on the parameters. They argue that this implicit prior is a reasonable one.

Kass and Wasserman (1995) note that the sample size n that appears in the penalty term of (1) must be carefully chosen. In censored data models such as the proportional hazards

model of Cox (1972), subjects contribute different amounts of information to the likelihood, depending on whether or not they are censored. We have found that, in the penalty term for BIC, substituting d, the number of uncensored events, for n, the total number of individuals, results in an improved criterion without sacrificing the asymptotic properties shown by Kass and Wasserman (1995).

The example in this paper, the Cardiovascular Health Study, is an observational study investigating 23 possible risk factors for cardiovascular disease in the older population. By using our revision, we get two main benefits: a better evaluation of the individual risk factor's association with the outcome and a better way of predicting who is most susceptible to disease in the future. Since the risk factor data are noninvasive and easily collected, our model can help provide an improved risk assessment model for cardiovascular disease.

The rest of the paper is organized as follows. Section 2 discusses Bayes factors and the Bayesian Information Criterion. Section 3 proposes our revision of BIC and discusses it in the context of two censored survival models, exponential regression and the Cox model. We analyze the Cardiovascular Health Study data in Section 4, comparing our revision to the standard method, and we provide a final discussion in Section 5.

2. The Bayesian Information Criterion

Standard Bayesian testing procedures use the Bayes factor (BF), which is the ratio of integrated likelihoods for two competing models. Kass and Raftery (1995) derive BIC as an approximation to twice the difference in log integrated likelihoods, so the difference in BIC between two models approximates twice the logarithm of the Bayes factor. Hence,

$$\frac{2\log(BF) - BIC}{2\log(BF)} \to 0. \tag{2}$$

However,

$$2\log(BF) - BIC \not\to 0. \tag{3}$$

Equation (3) implies that, for general priors on the parameters, $2\log(BF)$ − BIC has a nonvanishing asymptotic error of constant order, i.e., of order $O(1)$. Since the absolute value of BIC increases with n, the error tends to zero as a proportion of BIC. Therefore, BIC has the undesirable property that, for any constant k, BIC + k also approximates twice the log Bayes factor to the same order of approximation as BIC itself. This $O(1)$ error suggests that the BIC approximation is somewhat crude and may perform poorly for small samples.

Kass and Wasserman (1995) show that with nested models, under a particular prior on the parameters, the constant order asymptotic error disappears, and they argue that this prior can reasonably be used for inference purposes. Following the notation of their paper, let $Y = (y_1, \ldots, y_n)$ be i.i.d. observations from a family parameterized by (θ, ψ), with $\dim(\theta, \psi) = m$ and $\dim(\theta) = m_0$. Our goal is to test H_0: $\psi = \psi_0$ against H_1: $\psi \in \Re^{m-m_0}$ using the Bayes factor,

$$\mathrm{BF} = \frac{P(Y \mid H_0)}{P(Y \mid H_1)}. \tag{4}$$

The Bayesian information criterion (BIC) for testing H_0 versus H_1 is

$$\mathrm{BIC} = -2\left\{\ell_1(\hat{\theta}, \hat{\psi}) - \ell_0(\hat{\theta}_0)\right\} + (m - m_0)\log(n). \tag{5}$$

Let $I(\theta, \psi)$ be the $m \times m$ Fisher information matrix of (θ, ψ) associated with the full model, let $I_{\theta\psi}(\theta, \psi_0)$ denote the information matrix $(-E[(\partial^2\ell(\theta, \psi))/(\partial\theta\partial\psi)])$ evaluated at (θ, ψ_0), and let $\pi_\psi(\psi)$ be the marginal prior density of ψ under H_1. The main result of Kass and Wasserman (1995) is as follows. If the conditions that

(C1) the parameters are null orthogonal, i.e., $I_{\theta\psi}(\theta, \psi_0) = 0$ for all θ,
(C2) the MLE $\hat{\psi}$ satisfies $\hat{\psi} - \psi_0 = O_p(n^{-1/2})$, and
(C3) $-(1/n)D^2 l(\hat{\theta}, \hat{\psi}) - I(\theta, \psi) = O_p(n^{-1/2})$

hold, then

$$2\log \mathrm{BF} = \mathrm{BIC} - 2\log\left\{(2\pi)^{(m_0-m)/2} \times \left|I_{\psi\psi}(\hat{\theta}, \psi_0)\right|^{-1/2} \pi_\psi(\hat{\psi})\right\} + O_p\left(n^{-1/2}\right). \tag{6}$$

In addition, if $\pi_\psi(\psi)$ is a standard multivariate normal density with location ψ_0 and variance matrix $|I_{\psi\psi}(\theta, \psi_0)|^{-1}$, the asymptotic error of constant order (the second term on the right in (6)) will vanish, leaving

$$2\log(\mathrm{BF}) = \mathrm{BIC} + O_p\left(n^{-1/2}\right). \tag{7}$$

If the prior on ψ is not of this form, then this error term in (6) gives the constant order asymptotic error in BIC as an approximation to twice the log Bayes factor.

This result has an important implication. BIC is a Bayesian procedure that does not require the specification of a prior, but it approximates a Bayes factor that is based on a particular prior for the parameter of interest. Therefore, when using BIC to compare models, the Kass–Wasserman result defines an implicit prior, which BIC uses. This prior, which we call the overall unit-information prior, is appealing because it is a normal distribution centered around ψ_0 with the amount of information in the prior equal to the average amount of information in one observation. Since the prior is based on only one observation, it is vague yet proper.

3. Model Selection in Censored Survival Models

3.1 *Model Selection*

Model selection criteria such as BIC are often used to select variables in regression problems. Here, we use BIC to determine the best models (where models are variable subsets) in a class of censored survival models.

When censoring is present, it is unclear whether the penalty in BIC should use n, the number of observations, or d, the number of events. When using the partial likelihood (Cox, 1972), there are only as many terms in the partial likelihood as there are events d. Kass and Wasserman (1995) indicate that the term used in the penalty should be the rate at which the Hessian matrix of the log-likelihood function grows, which suggests that d is the correct quantity to use. However, if we are to use a revised version of BIC, it is important that the new criterion continue to have the asymptotic properties that Kass and Wasserman derived. In fact, our revised BIC does have these properties, with a slightly modified outcome. Let us alter condition (C3) to be

$$-\frac{1}{d}D^2 l\left(\hat{\theta},\hat{\psi}\right) - I_u(\theta,\psi) = O_p\left(n^{-1/2}\right), \qquad (8)$$

where $I_u(\theta,\psi)$ is the expected Fisher information for one uncensored observation, (the uncensored unit information). If this holds, then (7) is true, and the new BIC (with d in the penalty) is an $O_p(n^{-1/2})$ approximation to twice the Bayes factor, where the prior variance on θ is now equal to the inverse of the uncensored unit information. By using d in the penalty instead of n, we will show that this asymptotic result holds, the only difference being in the implicit prior on the parameter.

To investigate which of these criteria is better, we first consider a simple exponential censored survival model. We use a conjugate unit-information prior for this model and find that BIC provides a closer approximation to twice the log Bayes factor when n is replaced by d in equation (1). Then we consider the Cox proportional hazards model. The theorems stated in this section are proved in Volinsky and Raftery (1999).

3.2 *Exponential Survival*

Consider n subjects, where subject i has observed time y_i, and let $Y = (y_1, \ldots, y_n)$. The censoring indicator δ_i describes whether y_i is a survival time ($\delta_i = 1$) or a censoring time ($\delta_i = 0$). We consider testing H_0: $T_i \sim \mathrm{Exp}(\lambda_0)$ against H_1: $T_i \sim \mathrm{Exp}(\lambda)$; $\lambda \sim \mathrm{Gamma}(a,b)$, where T_i is the actual time of death for the ith subject (which will be unobserved if the observed time is a censoring time). BIC corresponds to an implicit normal prior, but the conjugate gamma prior in H_1 seems more appropriate here because it puts mass only on positive values of λ, while the normal prior puts some mass on negative values of λ, which should be excluded. We are interested in how well BIC approximates twice the log Bayes factor (equation (4)), where

$$P(Y \mid \mathrm{H}_i) = \int P(Y \mid \mathrm{H}_i, \lambda) P(\lambda \mid \mathrm{H}_i) d\lambda. \qquad (9)$$

Since the prior on λ is not normal, the BIC approximation to twice the log Bayes factor will have asymptotic error of constant order. We consider two unit-information Gamma priors, one corresponding to using n in the penalty and the other to using d. These priors have mean equal to the null hypothetical value (λ_0) and variance equal to the inverse expected information in one observation (for n) or in one uncensored observation (for d). In the n case, this yields a $\mathrm{Gamma}(q, q/\lambda_0)$ distribution, where q is the proportion of the data that is censored. In the d case, this yields a $\mathrm{Gamma}(1, 1/\lambda_0)$ distribution. We call these the overall unit-information prior and the uncensored unit-information prior, respectively. Using these priors, we can calculate the asymptotic difference between $2\log \mathrm{BF}$ and BIC for both n and d. We denote this asymptotic error by AE_n or AE_d, depending on which penalty term is used in BIC. The following theorem shows that, asymptotically, using d in the penalty term gives a closer approximation to twice the log Bayes factor.

THEOREM 1: *Consider comparing the exponential survival models* H_0 *and* H_1 *using BIC under independent censoring. Then* $|AE_d| < |AE_n|$.

In the context of epidemiological studies, this means that using BIC with d in the penalty will provide a better approximation to the Bayes factor for comparing two competing models.

3.3 *The Cox Proportional Hazards Model*

In the Cox (1972) proportional hazards model, estimation of β is commonly based on the partial likelihood, namely

$$PL(\beta) = \prod_{i=1}^{n} \left(\frac{\exp\left(\mathbf{x}_i^T \beta\right)}{\sum_{\ell \in R_i} \exp\left(\mathbf{x}_\ell^T \beta\right)} \right)^{w_i}, \qquad (10)$$

where R_i is the set of individuals at risk at time t_i, $\mathbf{x}_i$ is the vector of covariates for the ith individual, and w_i is a censoring indicator (uncensored = 1).

Suppose that we have survival data on n individuals with independent survival times $y_1, \ldots, y_n$ and independent censoring indicators $\delta_1, \ldots, \delta_n$. Consider the models H_1: $h_i(t) = \lambda_0(t)\exp(\mathbf{x}_i^T\beta)$ and H_0: $h_i(t) = \lambda_0^{(0)}(t)\exp(\mathbf{x}_i^{(0)T} \times \beta^{(0)})$, where $\beta^{(0)}$ is a subset of β and $\mathbf{x}_i^{(0)}$ is the corresponding subset of $\mathbf{x}_i$. Then H_0 and H_1 are nested models, with $\theta = \beta^{(0)}$, $(\theta,\psi) = \beta$, $m_0 = \dim(\beta^{(0)})$, and $m = \dim(\beta)$ in our previous notation. The following theorems show that the conclusions from Kass and Wasserman (1995) hold for Cox proportional hazards models whether n or d is used in the BIC penalty.

THEOREM 2: *Under null orthogonality and independent censoring, equation* (6) *holds for Cox models when* n *is used in the penalty for BIC and the normal overall unit-information prior is used.*

THEOREM 3: *Under null orthogonality and independent censoring, equation* (7) *holds for Cox models when* d *is used in the penalty for BIC and normal uncensored unit-information prior is used.*

It follows that (7) holds for the Cox model and that the $O(n^{-1/2})$ result also holds for both the overall and the uncensored unit-information priors with the appropriate penalty.

3.4 *Evaluation of the Unit Information Priors*

The choice of d or n in BIC determines the implicit prior variance. A desirable reference prior has most of its mass concentrated in that region of parameter space in which parameter estimates tend to fall, yet it is rather flat over that area. We looked at three commonly cited survival datasets, including the Cardiovascular Health Study and two others analyzing lung cancer (Prentice, 1973; Kalbfleisch and Prentice, 1980) and liver disease (Dickson, Fleming, and Weisner, 1985; Fleming and Harrington, 1991). We compared the priors for each choice of penalty to the parameter estimates eventually calculated.

For each of the studies, we found both the overall and the uncensored unit-information priors. These three datasets include a total of 43 candidate independent variables for predicting the risk of onset of a disease. Table 1 compares the unit-information priors for the three datasets. For clarity, the results have been standardized by the uncensored unit-information prior standard deviation. In the Cardiovascular

Table 1
Comparison of mean implicit prior standard deviation for the two different penalties in BIC for three censored survival data sets

Data	m	n	% Censored	Prior SD (d)	Prior SD (n)	Range of $\hat{\beta}_i$
CHS	23	4504	96	1.0	5.0	$(-0.09, 0.34)$
PBC	12	312	60	1.0	1.8	$(-0.29, 0.34)$
V.A.	8	137	7	1.0	1.0	$(-0.65, 0.43)$

Health Study, the overall unit-information prior standard deviations are, on average, five times greater than the average of the uncensored unit-information prior standard deviations, reflecting the large amount of censoring in these data. The parameter estimates (measured in units of one uncensored unit information prior standard deviation) range from -0.09 to $+0.34$, indicating that a prior standard deviation of 5.0 is much more spread out than necessary. For the three datasets, the range of all the estimates is $(-0.65, +0.43)$, indicating that the uncensored unit information prior (with a prior standard deviation of one on this scale) is more realistic because it covers the distribution of estimates very well without being much more spread out.

Figure 1 is a histogram of the estimated coefficients in the three datasets, with overall and uncensored unit-information priors also shown. The plot shows that the uncensored unit-information prior puts more prior mass on what is more likely to occur yet is still a rather conservative prior that allows for outlying estimates. It thus seems more satisfactory in practice for these data sets.

4. Example

4.1 *The Cardiovascular Health Study*

In this section, we analyze the Cardiovascular Health Study (CHS) in more detail to show the effect that changing the penalty term can have on the analysis as well as on predictive performance. The CHS is a longitudinal, observational study funded by the National Institutes of Health (NIH) and started in June 1989 to study cardiovascular disease in people aged 65 and over. The CHS studies 23 possible risk factors for stroke on 4504 subjects, 172 of whom did suffer a stroke during the study. One of the goals of the CHS researchers is to design a risk assessment model from these variables. To this end, all 23 variables can be collected via a patient's medical history or a noninvasive exam. An effective model built on these 23 variables can provide an efficient assessment of one's stroke risk based on a short doctor's visit. Obviously, there is great interest in any methodology that can provide an improved predictive model.

For these data, d is approximately 4% of n, so the choice of the penalty term will have a substantial impact on the analysis. In what follows, we refer to the analyses that use n and d in the penalty term as the n-analysis and the d-analysis, respectively.

For the n-analysis, the best model according to BIC has seven variables (regular aspirin use, systolic blood pressure, creatinine level, diabetes, presence of atrial fibrillation, stenosis of the carotid artery, and a timed walk). When d is used, the penalty for additional parameters is smaller, and the best model now contains nine variables —the same as above plus age and diuretic use. The question of whether to include age in the final model is an interesting one. Although

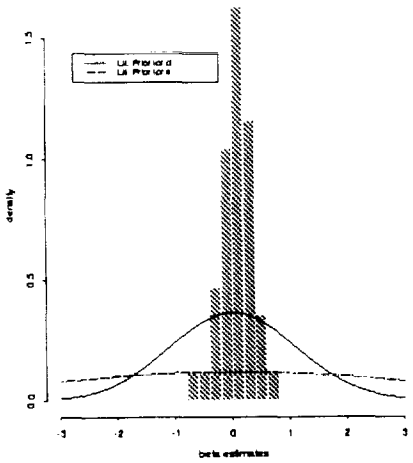

Figure 1. Histogram of standardized estimated coefficients from the three datasets in Table 1. A typical unit-information prior distribution corresponding to each of the penalty terms is shown.

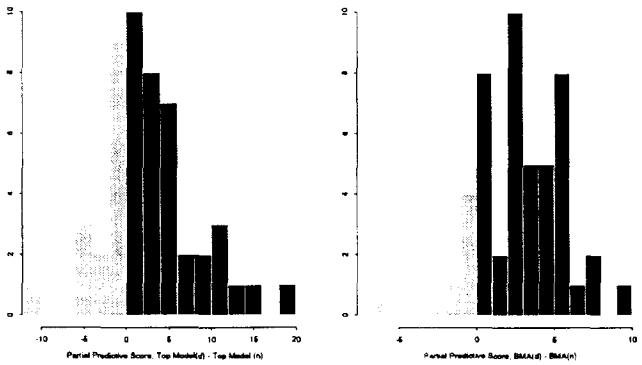

Figure 2. Histograms showing the difference in partial predictive score ($d - n$) on 100 splits of the data. A positive result (shown in black) indicates a win for d. On the left are the results of picking the single best model, on the right the results from model averaging.

Table 2
Posterior probabilities of parameters not being equal to zero and parameter estimates from Bayesian model averaging using n and d in the BIC penalty

Variable	$100\,P_n(\beta \neq 0)$	$100\,P_d(\beta \neq 0)$	$\hat{\beta}_n$	$\hat{\beta}_d$
Aspirin	88	100	0.48	0.52
Age	62	89	0.03	0.04
Atfib (ecg)	30	67	0.27	0.55
Diuretic use	27	59	0.12	0.23
LV hyper.	17	50	0.12	0.30

stroke incidence is known to increase with age, it is possible that age has no inherent effect on stroke beyond its effect on the other covariates. The d-analysis implies that age is indeed an important variable independent of the others. It may be acting as a surrogate for unidentified variables, or perhaps age itself is inherently a risk factor for stroke. Articles written in the medical literature on this dataset (Fried et al., 1991; Manolio et al., 1996) support the conclusion of the d-analysis that both age and diuretic use are important independent risk factors.

Another way of assessing the models is through predictive performance. A model with improved prediction will be able to provide a better risk evaluation of future patients. The data were split in half to create a model building set and a validation set. The models were assessed using an analogue of Good's log score (Good, 1952) called the partial predictive score (PPS) (Volinsky et al., 1997). For 50 different splits of the data, we calculated the difference in the PPS. Figure 2a shows the histogram of these 50 differences. In 35 of the 50 splits, the d-best model performed better than the n-best model, with an average of a 3.4% improvement in predictive performance per event. Similar improvements in predictive performance have been shown to make a substantive difference in risk classification models (Volinsky et al., 1997).

4.2 *Bayesian Model Averaging*

In variable selection problems, there is often model uncertainty due to the fact that the true model is not known. In our example, it is unclear whether age should be included in the model. It would be desirable to take account of this uncertainty instead of making a binary decision of in or out. One way of doing this is Bayesian model averaging (BMA). In a BMA analysis, the posterior distribution of any quantity of interest is a weighted average over all models, weighted by the support each model has from the data. So our analysis of age, or of risk scores, can take into account information from every model in the model class.

In several reported experiments with real data, BMA has yielded improved predictive performance and parameter estimation (Madigan and Raftery, 1994; Madigan, Gavrin, and Raftery, 1995; Raftery, Madigan, and Hoeting, 1997; Volinsky et al., 1997). An S-PLUS function to do Bayesian model averaging for Cox regression models, `bic.surv`, is available from the BMA homepage (`www.research.att.com/~volinsky/bma.html`) or by contacting the first author.

First, we specify a uniform prior over the space of the 2^{23} models associated with the candidate independent 23 variables in the CHS. This corresponds to an *a priori* probability of 50% on each individual variable. Then, after looking at the data, each model's posterior model probability (PMP) is calculated using BIC. To avoid fitting all possible models, we looked at only the best model and at all models with PMPs of at least 1/20 of the best model's PMP; this technique is known as Occam's Window (Madigan and Raftery, 1994). We used a branch and bound algorithm to find the models in Occam's Window (Volinsky et al., 1997). The d-analysis located 408 models ranging in size from 6 to 12 variables (median model size = 10), whereas the n-analysis found 168 models ranging in size from 5 to 9 variables (median model size = 7).

The BMA results for five of the variables are displayed in Table 2 and show how the two different penalties can lead to different conclusions. For instance, the coefficient of the variable aspirin has 100% posterior probability of being nonzero in the d-analysis, indicating very strong evidence that aspirin is a risk factor. In the n-analysis, by contrast, the aspirin parameter has a posterior probability of only 88%, which, by conventional rules of thumb (Kass and Raftery, 1995), is positive but not strong evidence for an effect. Several studies involving other data sets as well as this one (Steering Committee of the Physicians' Health Study, 1989; Stroke Prevention in Atrial Fibrillation Study Group Investigators, 1990; Paganini-Hill et al., 1989; Manolio et al., 1996) support the d-analysis conclusion of strongest possible evidence for aspirin as a risk factor for stroke, although they suggest that this risk may exist only for those with no history of cardiovascular disease. Manolio et al. (1996) conclude that the evidence for aspirin as a risk factor is too strong to ignore even though they do not understand the mechanism for the effect. The strength of the connection between aspirin and stroke clearly deserves further study.

For the next four variables, the data are inconclusive, but different conclusions would be drawn depending on the penalty. For instance, the posterior probability that the atrial fibrillation by ecg parameter is nonzero is 67% with the d prior and 30% with the n prior; this is the difference between weak evidence for an effect and evidence against an effect. In their expert analysis, Manolio et al. (1996) include all five of these variables in their multivariate model as significant or highly significant, supporting the d-analysis. Note that there are many other variables, including smoking, glucose level, and diastolic blood pressure, that get little support from both BMA analyses and Manolio et al. (1996).

Bayesian model averaging has been shown to provide a better risk assessment than selecting a single model (Volinsky et al., 1997). Using d in the penalty for BIC can provide a sharper improvement for BMA than using n, as shown in Figure 2b. For each of 50 splits of the data, we ran a BMA analysis using BIC to calculate posterior model probabilities and using the partial predictive score to assess predictive performance. In 42 of the 50 simulations, the BMA d-analysis predicted better than the BMA n-analysis. Overall, predictive performance per event improved by 3.1%.

5. Discussion

The unit-information prior is a reasonable one when there is little prior information. In fact, one could argue that there is always at least a small amount of prior information, namely that which led us to collect the data on the chosen independent variables in the first place. The unit-information prior quantifies this small bit of information as the average amount of information in one observation.

In this paper, we have contrasted two competing unit-information priors. The prior based on d is the unit-information prior for one uncensored observation (I_u) and as such does not depend on the proportion of censoring. In contrast, the overall unit-information prior (I) is typically qI_u (as in the exponential survival example) and will be affected by the proportion of censoring, which may not be desirable. To see this, consider two identical clinical trials where the first trial is stopped after 1 year and the second trial is stopped after 2 years. The second trial is likely to have a smaller percentage of censoring. The d-based prior for the two trials would be the same, apart from sampling variability. The n-based prior, in contrast, would be different; the n-based prior for the 1-year trial would be substantially more dispersed than that for the d-based trial. This strong dependence of the n-based prior on the stopping rule seems undesirable.

This methodology is not specific to censored survival models, but this is perhaps the most common example of models where data points may contribute different amounts of information to the likelihood. However, there are other models, such as weighted regression, measurement error, and missing data models, and our approach could also be used to find the correct penalty for BIC in these models.

Acknowledgements

This research was supported in part by ONR grant N00014-96-1-0192. The authors are grateful to two anonymous referees, whose comments led to improvements in the paper, and to Richard A. Kronmal for sharing the CHS data and for helpful discussions.

Résumé

Nous évaluons le critère d'information bayésien (BIC) pour la sélection de variables dans les modèles de survie pour données censurées. Kass et Wasserman (1995) ont montré que le BIC fournit une approximation proche du facteur de Bayes, lorsqu'on utilise une loi *a priori* d'information unitaire sur l'espace des paramètres. Nous proposons une révision du terme de pénalité du BIC, en le basant sur le nombre d'événements non-censurés, plutôt que sur le nombre d'observations. Pour un modèle à données censures simple, cette modification améliore l'approximation de facteur de Bayes exact basée sur une loi *a priori* d'information unitaire conjuguée. Dans le modèle des hasards proportionnels de Cox, nous proposons de définir BIC à partir de la vraisemblance partielle maximisée. L'utilisation dans le critère BIC du nombre de décès plutôt que du nombre d'individus correspond à une loi *a priori* sur l'espace des paramètres plus réaliste, et nous montrons qu'elle améliore la performance prédictive, pour l'estimation du risque d'infarctus dans une étude de santé cardio-vasculaire.

References

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

Dickson, E. R., Fleming, R. R., and Weisner, R. H. (1985). Trial of penicillamine in advanced primary biliary cirrhosis. *New England Journal of Medicine* **312,** 1011–1015.

Fleming, T. R. and Harrington, D. H. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Fried, L. P., Borhani, N. O., Enright, P., et al. (1991). The Cardiovascular Health Study: Design and rationale. *Annals of Epidemiology* **1,** 263–276.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B* **14,** 107–114.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses with large samples. *Journal of the American Statistical Association* **90,** 928–934.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. *Journal of the American Statistical Association* **89,** 1535–1546.

Madigan, D., Gavrin, J., and Raftery, A. E. (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Communications in Statistics—Theory and Methods* **24,** 2271–2292.

Manolio, T. A., Kronmal, R. A., Burke, G. L., et al. (1996). Short-term predictors of incident stroke in older adults. *Stroke* **27,** 1479–1486.

Paganini-Hill, A., Chao, A., Ross, R., and Henderson, B. (1989). Aspirin use and chronic diseases: A cohort study of the elderly. *British Medical Journal* **299,** 1247–1250.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika* **60,** 279–288.

Raftery, A. E. (1995). Bayesian model selection in social research (with discussion). In *Sociological Methodology,* P. V. Marsden (ed), 111–195. Cambridge, Massachusetts: Blackwell.

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

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6,** 461–464.

Steering Committee of the Physicians' Health Study. (1989). Final report on the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine* **321,** 129–135.

Stroke Prevention in Atrial Fibrillation Study Group Investigators. (1990). Preliminary report of the stroke prevention in atrial fibrillation. *New England Journal of Medicine* **322,** 863–868.

Volinsky, C. T. and Raftery, A. E. (1999). *Bayesian information criterion for censored survival models.* Technical Report 349, Department of Statistics, University of Washington, Seattle. (http://www.stat.washington.edu/tech.reports/tr349.ps).

Volinsky, C. T., Madigan, D., Raftery, A. E., and Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Applied Statistics* **46,** 443-448.

Received July 1998. *Revised June* 1999.
Accepted June 1999.