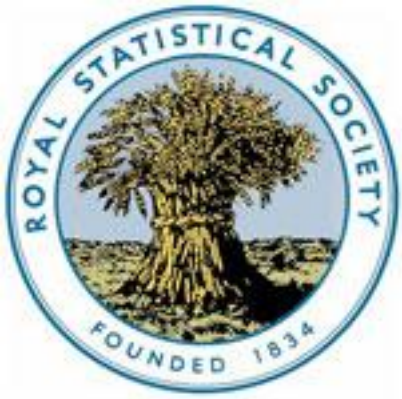

Prior Elicitation, Variable Selection and Bayesian Computation for Logistic Regression Models

Author(s): Ming-Hui Chen, Joseph G. Ibrahim and Constantin Yiannoutsos

Source: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 61, No. 1 (1999), pp. 223-242
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2680747
Accessed: 08-08-2016 05:51 UTC

J. R. Statist. Soc. B (1999)
61, *Part* 1, *pp.* 223–242

Prior elicitation, variable selection and Bayesian computation for logistic regression models

Ming-Hui Chen,

Worcester Polytechnic Institute, USA

Joseph G. Ibrahim

Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, USA

and Constantin Yiannoutsos

Harvard School of Public Health, Boston, USA

[Received August 1996. Final revision February 1998]

Summary. Bayesian selection of variables is often difficult to carry out because of the challenge in specifying prior distributions for the regression parameters for all possible models, specifying a prior distribution on the model space and computations. We address these three issues for the logistic regression model. For the first, we propose an informative prior distribution for variable selection. Several theoretical and computational properties of the prior are derived and illustrated with several examples. For the second, we propose a method for specifying an informative prior on the model space, and for the third we propose novel methods for computing the marginal distribution of the data. The new computational algorithms only require Gibbs samples from the full model to facilitate the computation of the prior and posterior model probabilities for all possible models. Several properties of the algorithms are also derived. The prior specification for the first challenge focuses on the observables in that the elicitation is based on a prior prediction y_0 for the response vector and a quantity a_0 quantifying the uncertainty in y_0. Then, y_0 and a_0 are used to specify a prior for the regression coefficients semi-automatically. Examples using real data are given to demonstrate the methodology.

Keywords: Gibbs sampling; Logistic regression; Normal prior; Posterior distribution; Prior distribution; Selection of variables

1. Introduction

The selection of variables is one of the most frequently encountered problems in statistical data analysis. In cancer or clinical trials for acquired immune deficiency syndrome (AIDS), for example, one often wishes to assess the importance of certain prognostic factors such as treatment, age, gender or race in predicting survival outcome. Most of the existing literature addresses variable selection for logistic regression by using criterion-based methods such as the Akaike information criterion (AIC) (Akaike, 1973) or the Bayes information criterion (BIC) (Schwarz, 1978). Fully Bayesian approaches to variable selection for these models is now feasible because of recent advances in computing technology and the development of efficient computational algorithms. There has been a recent surge in the development of

Address for correspondence: Joseph G. Ibrahim, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA.
E-mail: ibrahim@jimmy.harvard.edu

 1369–7412/99/61223

Bayesian methods for analysing logistic regression models. Papers which address informative prior specifications for the logistic regression model include Zellner and Rossi (1984), West (1985), West *et al.* (1985), Albert (1988), Zeger and Karim (1991), Albert and Chib (1993), Gelfand *et al.* (1996), Müller and Roeder (1996) and Bedrick *et al.* (1996). These papers mainly address the issue of Bayesian calculations for logistic regression when there is no uncertainty regarding the model itself. Other papers addressing informative prior elicitation and the selection of variables include George *et al.* (1996), who implemented an adaptation of their normal linear regression algorithm (see George and McCulloch (1993)), and Raftery (1996), who examined Bayes factors in accounting for model uncertainty.

In this paper, we examine the problem of eliciting informative prior distributions for the regression parameters as well as the elicitation of an informative prior distribution for the model space for Bayesian selection of variables in logistic regression. We propose a class of informative priors that appear to be quite useful in practice. Theoretical and computational properties of the priors are derived. Another major contribution of this paper is to develop novel computational algorithms for computing analytically intractable prior and posterior model probabilities. These algorithms are efficient and only require Gibbs samples from a single model. Theoretical properties of the algorithms are also derived. The Bayesian approach to variable selection is straightforward in principle. One quantifies the prior uncertainties via probabilities for each model under consideration, specifies a prior distribution for each of the parameters in each model and then uses Bayes theorem to calculate posterior model probabilities. In addition to the computational issues for calculating posterior probabilities (see Section 3), there are other difficulties in carrying out this procedure. Specifying meaningful prior distributions for the parameters in each model is a difficult task requiring contextual interpretations of a large number of parameters. A need arises then to look for some useful automated specifications. Reference priors can be used in many situations to address this. In some cases, however, they lead to ambiguous posterior probabilities and require problem-specific modifications such as those in Smith and Spiegelhalter (1980). Recently, Berger and Pericchi (1996) have proposed a set of measures which they call 'intrinsic Bayes factors' that provide a generic solution to the ambiguity problem. However, reference priors exclude the use of any real prior information that we may have. Even if we overcome the problem of specifying priors for the parameters in the various models, there remains the question of choosing prior probabilities for the models themselves. Here, the default prior giving equal probabilities to all models under consideration may not be satisfactory for certain problems, as demonstrated in Section 4. To overcome such difficulties, Ibrahim and Laud (1994) and Laud and Ibrahim (1995) advocated an approach to the selection of variables for the linear model by adopting the philosophy in Geisser (1993). The approach they recommend for the linear model is based on the notion of specifying a prior prediction y_0 for the response vector, and a scalar a_0 which quantifies one's assignment of information contributing to this guess relative to the information to be collected in the experiment. Then, y_0 and a_0, along with the design matrix for model m, are used as prior information to specify an automated parametric informative prior for the regression coefficients $\beta^{(m)}$. The motivation behind this approach is that the investigator often has prior information from similar past studies measuring the same response variable and covariates as for the current study.

The methodology proposed here is well suited for cancer and AIDS clinical trials research. To motivate the methodology, we consider the two landmark AIDS clinical trials ACTG019 and ACTG036. ACTG019 was an AIDS clinical trial comparing zidovudine (AZT) with a placebo. The response variable was binary with 1 denoting death, AIDS or AIDS-related

complex (ARC) and 0 otherwise. The results of this study were published in 1990 in the *New England Journal of Medicine*. In the ACTG019 study several important prognostic factors were identified for predicting the response. Among these were age, treatment and CD4 cell count. The clinical trial ACTG036, whose results were published a year later, was an almost identical trial also comparing AZT with a placebo. The ACTG036 trial had the same response variable and had many covariates in common with the ACTG019 study. These two trials fit into the setting considered here. We can use the covariates age, treatment and CD4 cell count from the ACTG019 trial as prior input to the ACTG036 trial. Here, y_0 can be taken as the response vector from the ACTG019 trial, along with its corresponding covariate matrix $X_0^{(m)}$ under model m.

The remainder of the paper is organized as follows. In Section 2.2 we propose a family of informative prior distributions for the regression parameters and investigate its properties. In Section 2.3, we propose a prior distribution for the model space, and in Section 3 we present a novel Monte Carlo method for estimating the marginal distribution of the data, and we provide a Markov chain Monte Carlo sampling scheme that will facilitate the computation of prior and posterior model probabilities. In Section 4, we conduct a simulation study and examine the ACTG019 and ACTG036 clinical trials in more detail. We conclude with a discussion section.

2. The method

2.1. Model and notation

Let k denote the number of covariates for the full model and let $\mathcal{M}$ denote the model space. We enumerate the models in $\mathcal{M}$ by $m = 1, 2, \ldots, \mathcal{K}$, where $\mathcal{K}$ is the dimension of $\mathcal{M}$ and model $\mathcal{K}$ denotes the full model. Also, let $\beta^{(\mathcal{K})} = (\beta_0, \beta_1, \ldots, \beta_k)'$ denote the regression coefficients for the full model including an intercept, and let $\beta^{(m)}$ denote a $k_m \times 1$ vector of regression coefficients for model m with an intercept, and a specific choice of $k_m - 1$ covariates. We write $\beta^{(\mathcal{K})} = (\beta^{(m)\prime}, \beta^{(-m)\prime})'$, where $\beta^{(-m)}$ is $\beta^{(\mathcal{K})}$ with $\beta^{(m)}$ deleted. Under model m, the likelihood function based on n observations for the current study is given by

$$L(\beta^{(m)}|D^{(m)}) = \exp(y'X^{(m)}\beta^{(m)} - J'Q^{(m)}), \tag{2.1}$$

where $y = (y_1, \ldots, y_n)'$ denotes the $n \times 1$ vector of binary responses, J is an $n \times 1$ vector of 1s and $X^{(m)}$ is an $n \times k_m$ matrix of fixed covariates of rank k_m. Also, $Q^{(m)}$ is an $n \times 1$ vector with jth element $\log\{1 + \exp(x_j^{(m)\prime}\beta^{(m)})\}$, where $x_j^{(m)\prime}$ denotes the jth row of $X^{(m)}$. Finally, $D^{(m)} = (n, y, X^{(m)})$ denotes the data for the current study under model m.

2.2. The prior distributions

Our prior construction is based on the notion of the existence of a previous study that measures the same response variable and covariates as the current study. For ease of exposition, we assume only one previous study, as the extension to multiple previous studies is straightforward. For this, let n_0 denote the sample size for the previous study, y_0 be an $n_0 \times 1$ response vector for the previous study, $X_0^{(m)}$ be an $n_0 \times k_m$ matrix of covariates corresponding to y_0 and $D_0^{(m)} \equiv (n_0, y_0, X_0^{(m)})$ denote the prior data for model m. We shall often refer to $D_0^{(m)}$ as the historical data throughout. Further, let $\pi_0(\beta^{(m)}|\cdot)$ denote the prior distribution for $\beta^{(m)}$ from the previous study. Using this information, we wish to construct a prior distribution for $\beta^{(m)}$ for the current study. We propose a prior of the form

$$\pi(\beta^{(m)}|D_0^{(m)}, a_0) \propto \exp\{a_0(y_0'X_0^{(m)}\beta^{(m)} - J_0'Q_0^{(m)})\}\, \pi_0(\beta^{(m)}|c_0), \tag{2.2}$$

where J_0 is an $n_0 \times 1$ vector of 1s, $Q_0^{(m)}$ is an $n_0 \times 1$ vector with jth component $\log\{1 + \exp(x_{j0}^{(m)'}\beta^{(m)})\}$, c_0 is a fixed hyperparameter, a_0 is a scalar prior parameter that weights the prior data relative to the likelihood of the current study and $\pi_0(\beta^{(m)}|c_0)$ is the *initial prior* for $\beta^{(m)}$, i.e. $\pi_0(\beta^{(m)}|c_0)$ is the prior for $\beta^{(m)}$ for the previous study. The prior parameter c_0 controls the effect of $\pi_0(\beta^{(m)}|c_0)$ on the entire prior, and the parameter a_0 controls the influence of the prior data on $\pi(\beta^{(m)}|D_0^{(m)}, a_0)$. The parameter a_0 can be interpreted as a dispersion parameter for the prior data. It is reasonable to restrict the range of a_0 to be between 0 and 1, and thus we take $0 \leqslant a_0 \leqslant 1$. One of the main roles of a_0 is that it controls the heaviness of the tails of the prior for $\beta^{(m)}$. As a_0 becomes smaller, the tails of distribution (2.2) become heavier.

The most natural specification of $D_0^{(m)}$ is to take y_0 to be the raw response vector from the previous study, to take $X_0^{(m)}$ to be the raw covariate matrix under model m from the previous study and to take n_0 to be the sample size of the previous study. In this case expression (2.2) has several appealing interpretations. The first term on the right-hand side is just the likelihood function of $\beta^{(m)}$ based on the historical data $D_0^{(m)}$ raised to the power a_0. Setting $a_0 = 1$, expression (2.2) corresponds to the update of $\pi_0(\beta^{(m)}|c_0)$ using Bayes theorem, i.e. with $a_0 = 1$ it corresponds to the posterior distribution of $\beta^{(m)}$ from the previous study. When $a = 0$, the prior does not depend on the historical data, and in this case $\pi(\beta^{(m)}|D_0^{(m)}, a_0) \equiv \pi_0(\beta^{(m)}|c_0)$. Therefore, prior (2.2) can be viewed as a generalization of the usual Bayesian update of $\pi_0(\beta^{(m)}|c_0)$. The parameter a_0 allows the investigator to control the influence of the historical data on the current study. Such control is important in cases where there is heterogeneity between the previous and current study, or when the sample sizes of the two studies are quite different.

The prior specification is completed by specifying a prior distribution for a_0. We take a beta prior for a_0, and thus we propose a joint prior distribution for $(\beta^{(m)}, a_0)$ of the form

$$\pi(\beta^{(m)}, a_0|D_0^{(m)}) \propto \exp\{a_0(y_0'X_0^{(m)}\beta^{(m)} - J_0'Q_0^{(m)})\}\, a_0^{\delta_0 - 1}(1 - a_0)^{\lambda_0 - 1}\, \pi_0(\beta^{(m)}|c_0), \tag{2.3}$$

where (δ_0, λ_0) are specified prior parameters. The prior in expression (2.3) does not have a closed form but it has several attractive theoretical and computational properties. First, we note that, if $\pi_0(\beta^{(m)}|c_0)$ is proper, then prior (2.3) is guaranteed to be proper. Further, prior (2.3) can be proper even if $\pi_0(\beta^{(m)}|c_0)$ is improper. Chen *et al.* (1997) gave sufficient conditions for the propriety of prior (2.3) in the case that $\pi_0(\beta^{(m)}|c_0)$ is a uniform improper prior and also showed that prior (2.2) converges to a multivariate normal distribution as $n_0 \to \infty$. We refer the reader to Chen *et al.* (1997) for more details on the theoretical properties of the priors. One attractive feature of expression (2.3) is that it creates heavier tails for the marginal prior of $\beta^{(m)}$ than does prior (2.2), which assumes that a_0 is a fixed value. This is a desirable feature since it gives the investigator more flexibility in weighting the historical data. In addition, our construction of expression (2.3) is quite general, with various possibilities for $\pi_0(\beta^{(m)}|c_0)$. Specific choices are discussed in Section 2.4.

If a previous study does not exist on which to base $D_0^{(m)}$, then y_0 can be obtained via a prior prediction, including specifications based on a theoretical prediction model, expert opinion or case-specific information. For example, a theoretical model of the form $y_0 = g(X^{(m_0)})$ may be available for obtaining the prior predictions, where $X^{(m_0)}$ is the covariate matrix corresponding to some model m_0 and g is a known function. Such prediction models are often used, for example, in respiratory studies measuring forced vital capacity and forced expiratory volume. Also, in these cases, we may take $X_0^{(m)}$ to be the covariate matrix of the current study,

i.e. $X_0^{(m)} = X^{(m)}$ and $n_0 = n$. In any case, the existence of a previous study leads to the most natural specification of $D_0^{(m)}$ and serves as the primary motivation for expression (2.3). Taking $D_0^{(m)}$ to be the raw data from a previous study results in a more natural, interpretable and automated specification for expression (2.3).

2.3. *A generalization of the priors*

When a previous study is available, it sometimes occurs that the set of covariates measured in the previous study is a subset of the covariates measured in the current study. This may occur because the investigators discover 'new' and potentially useful covariates to measure in the current study that were not measured in previous studies. In this case, we can modify expression (2.2) as follows. Let $X_1^{(m)}$ denote the $n \times r_m$ matrix of covariates in the current study that are common to the covariates in the previous study, and let $X_2^{(m)}$ be the $n \times s_m$ matrix of new covariates in the current study which are not measured in the previous study. Write

$$\beta^{(m)} = \begin{pmatrix} \beta_1^{(m)} \\ \beta_2^{(m)} \end{pmatrix},$$

and let $X_{01}^{(m)}$ represent the $n_{01} \times r_m$ matrix of covariates from the previous study; $X_{02}^{(m)}$ is an $n_{02} \times s_m$ matrix of covariates representing the new covariates and $k_m = r_m + s_m$. The most natural choice for $X_{01}^{(m)}$ is the raw covariate matrix from the previous study, and to take $X_{02}^{(m)} = X_2^{(m)}$. In our prior specification, we assume that the new covariates have small or negligible correlation with the common covariates, i.e. $\text{corr}(X_1^{(m)}, X_2^{(m)}) \approx 0$. This seems to be a sensible assumption if the new set of covariates in the current study is being scientifically investigated for the first time. Finally, we assume *a priori* independence between $\beta_1^{(m)}$ and $\beta_2^{(m)}$, which leads to

$$\begin{aligned} \pi(\beta^{(m)}|D_0^{(m)}, a_0) &= \pi_1(\beta_1^{(m)}|D_{01}^{(m)}, a_{01})\, \pi_2(\beta_2^{(m)}|D_{02}^{(m)}, a_{02}) \\ &\propto \exp\{a_{01}(y_{01}' X_{01}^{(m)} \beta_1^{(m)} - J_{01}' Q_{01}^{(m)})\} \\ &\quad \times \exp\{a_{02}(y_{02}' X_{02}^{(m)} \beta_2^{(m)} - J_{02}' Q_{02}^{(m)})\}\, \pi_0(\beta_1^{(m)}, \beta_2^{(m)}|c_0), \end{aligned} \tag{2.4}$$

where y_{01} and y_{02} represent vectors of prior predictions, a_{0j} is a prior parameter and $D_{0j}^{(m)} = (n_{0j}, y_{0j}, X_{0j}^{(m)})$, $j = 1, 2$. The prior specification is completed by specifying independent beta priors for (a_{01}, a_{02}), leading to the joint prior

$$\pi(\beta_1^{(m)}, \beta_2^{(m)}, a_{01}, a_{02}) \propto \prod_{j=1}^{2} [\exp\{a_{0j}(y_{0j}' X_{0j}^{(m)} \beta_j^{(m)} - J_{0j}' Q_{0j}^{(m)})\}\, a_{0j}^{\delta_{0j}} (1 - a_{0j})^{\lambda_{0j}-1}]\, \pi_0(\beta_1^{(m)}, \beta_2^{(m)}|c_0). \tag{2.5}$$

A natural choice for y_{01} is the raw response vector from the previous study. The elicitation of y_{02} is less automatic since no *a priori* information is available for it. One possible choice is to pick $y_{02} = (\frac{1}{2}, \ldots, \frac{1}{2})$. This choice results in $\pi(\beta_2^{(m)}|D_{02}^{(m)}, a_{02})$ having a mode equal to 0. Also we take

$$\pi_0(\beta_1^{(m)}, \beta_2^{(m)}|c_0) = \pi_0(\beta_1^{(m)}|c_0)\, \pi_0(\beta_2^{(m)}|c_0).$$

The prior parameters $(\delta_{02}, \lambda_{02})$ are chosen to reflect vague prior beliefs, and thus values such

as $\delta_{02} = \lambda_{02} = 1$ (i.e. a uniform prior) would be reasonable. We mention that we cannot take $a_{02} = 0$ with probability 1 since this would make the prior improper when a flat initial prior for $\beta_2^{(m)}$ is used, i.e. $\pi_0(\beta_2^{(m)}|c_0) \propto 1$, and thus we would not be able to compute posterior model probabilities when no information is available to specify the initial prior $\pi_0(\beta_2^{(m)}|c_0)$ for $\beta_2^{(m)}$. Prior (2.5) reduces to prior (2.3) if the sets of covariates from the previous and current studies are identical. If the set of covariates in the current study is a subset of the covariates in the previous study, then we can construct a submatrix by omitting those columns corresponding to covariates that are not in the current study and we take $X_0^{(m)}$ to be that submatrix.

2.4. Choices of prior parameters

There are several ways in which we can choose the prior parameters (δ_0, λ_0). For elicitation, it is often easier to work with $\mu_0 = \delta_0/(\delta_0 + \lambda_0)$ and

$$\sigma_0^2 = \mu_0(1 - \mu_0)(\delta_0 + \lambda_0 + 1)^{-1}.$$

A uniform prior (i.e. $\delta_0 = \lambda_0 = 1$), which corresponds to $(\mu_0, \sigma_0^2) = (1/2, 1/12)$, may be a suitable non-informative starting point and facilitates a useful reference analysis for other choices. The investigator may choose μ_0 to be small (say $\mu_0 \leqslant 0.1$), if he or she wishes to have low prior weight on the historical data. If a large prior weight is desired, then $\mu_0 \geqslant 0.5$ may be desirable. It is reasonable to choose σ_0^2 in the range $\mu_0/1000 \leqslant \sigma_0^2 \leqslant \mu_0/10$. For the generalized prior in expression (2.5), the prior parameters can be chosen in a similar manner. Also, it is desirable in this case to take $\mu_{02} \leqslant \mu_{01}$, where $\mu_{0j} = \delta_{0j}/(\delta_{0j} + \lambda_{0j})$. Choices of the form $\mu_{02} = p_0\mu_{01}$, where $0 \leqslant p_0 \leqslant 1$, are suitable. In any case, in an actual analysis, we recommend that several choices of (μ_0, σ_0^2) be used, including choices that give small and large weight to the historical data, and that several sensitivity analyses be conducted. We do not recommend doing an analysis based on one set of prior parameters. The choices recommended here can be used as starting points from which sensitivity analyses can be based.

It is reasonable to specify a non-informative prior for $\pi_0(\beta^{(m)}|c_0)$ since this is the prior for $\beta^{(m)}$ corresponding to the previous study and contains no information about the historical data $D_0^{(m)}$. For this, one choice is to take $\pi_0(\beta^{(m)}|c_0)$ to be a normal density with mean 0 and covariance matrix $c_0 W_0^{(m)}$, i.e.

$$\pi_0(\beta^{(m)}|c_0) = (2\pi)^{-k_m/2} c_0^{-k_m/2} |W_0^{(m)}|^{-1/2} \exp\left\{ -\frac{1}{2c_0} \beta^{(m)\prime} (W_0^{(m)})^{-1} \beta^{(m)} \right\}. \tag{2.6}$$

The quantity $c_0 \geqslant 0$ is a scalar dispersion parameter which controls the effect of $\pi_0(\beta^{(m)}|c_0)$ on $\pi(\beta^{(m)}, a_0|D_0^{(m)})$ and hence influences the marginal distribution of the data. To make $\pi_0(\beta^{(m)}|c_0)$ non-informative, we take large values of c_0 so that $\pi_0(\beta^{(m)}|c_0)$ is flat relative to the other terms in expression (2.3). Small values of c_0 will let $\pi_0(\beta^{(m)}|c_0)$ dominate expression (2.3). Thus, c_0 is an important tuning parameter that allows us to control the effect of the marginal distribution of the data for the calculation of posterior model probabilities.

The size of c_0 that is used will depend on the structure of the data set and the prior parameters for a_0. From the examples of Section 4, reasonable choices of c_0 are $c_0 \geqslant 5$. In any case, we do not recommend an automatic one-time specification for c_0, but rather we emphasize that several sensitivity analyses should be conducted with several values of c_0 to examine the effect of $\pi_0(\beta^{(m)}|c_0)$ on the posterior model probabilities.

The matrix $W_0^{(m)}$ has a less crucial role than c_0 and is specified as follows. Let $W_0^{(K)}$ be a diagonal matrix with the ith diagonal element equal to the ith diagonal element of $(X_0^{(K)\prime} V_0^{(K)} X_0^{(K)})^{-1}$, where $V_0^{(K)}$ is the $n_0 \times n_0$ diagonal matrix with ith element

$$v_{0i}^{(\mathcal{K})} = \frac{\exp(x_{0i}^{(\mathcal{K})}\hat{\beta}_0)}{\{1+\exp(x_{0i}^{(\mathcal{K})}\hat{\beta}_0)\}^2}$$

and $\hat{\beta}_0$ is the maximum likelihood estimator of β based on the historical data $D_0^{(\mathcal{K})} = (n_0, y_0, X_0^{(\mathcal{K})})$ for the full model. Thus, the diagonal elements of $W_0^{(\mathcal{K})}$ correspond to the asymptotic variances of $\hat{\beta}_0$ based on the full model. Now we take $W_0^{(m)}$ to be the submatrix of the diagonal matrix $W_0^{(\mathcal{K})}$ corresponding to model m. The purpose of picking $W_0^{(m)}$ in this way is to adjust properly for the different scales of the measured covariates. If the covariates are all standardized or are measured on the same scale, then we take $W_0^{(m)} = I$. In any case, $W_0^{(m)}$ plays a minimal role when c_0 is taken large.

Another choice for $\pi_0(\beta^{(m)}|c_0)$ that we consider is a uniform improper prior, i.e. $\pi_0(\beta^{(m)}|c_0) \propto 1$. This corresponds to the case $c_0 \to \infty$ in equation (2.6) and thus can be viewed as a special case of it. As shown in Chen *et al.* (1997), under certain minor regularity conditions $\pi(\beta^{(m)}, a_0|D_0^{(m)})$ is a proper prior when $\pi_0(\beta^{(m)}|c_0) \propto 1$. This is a very important property since it tells us that prior (2.3) remains well defined as $c_0 \to \infty$ which is useful for Gibbs sampling from prior (2.3).

2.5. *Prior distribution on the model space*

Let

$$p_0^*(\beta^{(m)}|D_0^{(m)}) = \exp(y_0' X_0^{(m)}\beta^{(m)} - J_0' Q_0^{(m)})\,\pi_0(\beta^{(m)}|d_0), \tag{2.7}$$

where $\pi_0(\beta^{(m)}|d_0)$ is the same density as that described in Section 2.4 with c_0 replaced by d_0. We propose to take the prior probability of model m as

$$\begin{aligned} p(m) &\equiv p(m|D_0^{(m)}) \\ &= \frac{\int p_0^*(\beta^{(m)}|D_0^{(m)})\,\mathrm{d}\beta^{(m)}}{\sum_{m\in\mathcal{M}} \int p_0^*(\beta^{(m)}|D_0^{(m)})\,\mathrm{d}\beta^{(m)}}. \end{aligned} \tag{2.8}$$

The parameter d_0 is a scalar prior parameter that controls the effect of $\pi_0(\beta^{(m)}|d_0)$ on the prior model probability $p(m)$. This choice for $p(m)$ has several nice interpretations. First, $p(m)$ in equation (2.8) corresponds to the posterior probability of model m based on the data $D_0^{(m)}$ using a uniform prior for the previous study, $p_0(m) = 2^{-k}$ for $m \in \mathcal{M}$, i.e. $p(m) \propto p(m|D_0^{(m)})$, and thus $p(m)$ corresponds to the usual Bayesian update of $p_0(m)$ using $D_0^{(m)}$ as the data. Second, as $d_0 \to 0$, $p(m)$ reduces to a uniform prior on the model space. Therefore, as $d_0 \to 0$, the historical data $D_0^{(m)}$ have a minimal effect in determining $p(m)$. In contrast, with a large value of d_0, $\pi_0(\beta^{(m)}|d_0)$ plays a minimal role in determining $p(m)$, and in this case the historical data play a larger role in determining $p(m)$. Thus, as $d_0 \to \infty$, $p(m)$ will be regulated by the historical data. The parameter d_0 plays the same role as c_0 and thus serves as a tuning parameter to control the effect of $D_0^{(m)}$ on the prior model probability $p(m)$.

It is important to note that we use a scalar parameter c_0 in constructing the prior distribution $\pi(\beta^{(m)}, a_0|D_0^{(m)})$ given in equation (2.3), whereas we use a *different* scalar parameter d_0 in determining $p(m)$. This development provides us with great flexibility in specifying the prior distribution for $\beta^{(m)}$ as well as the prior model probabilities $p(m)$. In addition, as shown in Chen *et al.* (1997), if $\pi_0(\beta^{(m)}|d_0)$ is a uniform (improper) prior, then under mild conditions

$$\int \exp(y_0' X_0^{(m)} \beta^{(m)} - J_0' Q_0^{(m)})\, \mathrm{d}\beta^{(m)} < \infty,$$

and therefore formula (2.8) is well defined even if $\pi_0(\beta^{(m)}|d_0)$ is a uniform prior.

We mention here that the prior for $(\beta^{(m)}, a_0)$ in expression (2.3) does not imply a probability structure for $p(m)$ or any specific form for $p(m)$. Thus the prior for a_0 has only an effect on the marginal distribution of the data and has no effect whatsoever on $p(m)$. Thus, whether we have large or small weight on the historical data, it does not affect the numerical value of $p(m)$. Therefore $p(m)$ is the same regardless of the choice of prior parameters for a_0. In addition the prior for $(\beta^{(m)}, a_0)$ is not affected by the choice of $p(m)$, and in particular the choice of d_0. Finally, we note that if we let $d_0 \to 0$, $a_0 \to 0$ and $c_0 \to \infty$ then the posterior model probability is completely regulated by the likelihood and, in this case, would yield results that are similar to those of criterion-based likelihood methods such as the AIC and BIC.

3. Computation of model probabilities

In this section, we propose novel Monte Carlo implementation procedures to compute prior and posterior model probabilities. For the discussion, it suffices to consider expression (2.3), since the extension to expression (2.4) is clear.

3.1. Computing prior model probabilities

To compute $p(m)$ in equation (2.8), we adopt a Monte Carlo approach similar to Ibrahim *et al.* (1996) to estimate all the prior model probabilities by using a single Gibbs sample from the full model. The details of this procedure are given in Appendix A.

Following the Monte Carlo method of Chen and Shao (1997a) the prior probability of model m can be estimated by

$$\hat{p}(m) \equiv \hat{p}(m|D_0^{(m)}) = \frac{(1/N_0)\sum_{l=1}^{N_0} p_0^*(\beta_{0(l)}^{(m)}|D_0^{(m)})\, w(\beta_{0(l)}^{(-m)}|\beta_{0(l)}^{(m)})/p_0^*(\beta_{0(l)}^{(K)}|D_0^{(K)})}{(1/N_0)\sum_{j=1}^{K}\sum_{l=1}^{N_0} p_0^*(\beta_{0(l)}^{(j)}|D_0^{(j)})\, w(\beta_{0(l)}^{(-j)}|\beta_{0(l)}^{(j)})/p_0^*(\beta_{0(l)}^{(K)}|D_0^{(K)})}, \tag{3.1}$$

where $p_0^*(\beta^{(m)}|D_0^{(m)})$ is given by equation (2.7), $\beta_{0(l)}^{(K)} = (\beta_{0(l)}^{(m)\prime}, \beta_{0(l)}^{(-m)\prime})'$, $l = 1, 2, \ldots, N_0$, are samples from

$$p_0(\beta^{(K)}|D_0^{(K)}) \propto p_0^*(\beta^{(K)}|D_0^{(K)}) \tag{3.2}$$

and $w(\beta^{(-m)}|\beta^{(m)})$ is a *completely* known conditional density whose support is contained in or equal to the support of the conditional density of $\beta^{(-m)}$ given $\beta^{(m)}$ with respect to the full model joint prior distribution (3.2). The implementation of Markov chain Monte Carlo sampling, the justification for $\hat{p}(m)$, and the procedure to construct a good $w(\beta^{(-m)}|\beta^{(m)})$ are given in Appendix A.

There are several advantages of the above Monte Carlo procedure. Firstly, we need only one random draw from $p_0(\beta^{(K)}|D_0^{(K)})$, which greatly eases the computational burden. Secondly, it is more numerically stable since we calculate ratios of the densities in equation (3.1). Thirdly, in equation (3.1), $p_0(\beta^{(K)}|D_0^{(K)})$ plays the role of a ratio importance sampling density (see Chen and Shao (1997b)) which needs to be known only up to a normalizing constant since this common constant is cancelled out in the calculation.

3.2. Computing posterior model probabilities

The posterior probability of model m is given by

$$p(m|D^{(m)}) = p(D^{(m)}|m)\,p(m) \Big/ \sum_{m\in\mathcal{M}} p(D^{(m)}|m)\,p(m), \tag{3.3}$$

where

$$p(D^{(m)}|m) = \int L(\beta^{(m)}|D^{(m)})\,\pi(\beta^{(m)}, a_0|D_0^{(m)})\,\mathrm{d}\beta^{(m)}\,\mathrm{d}a_0 \tag{3.4}$$

denotes the marginal distribution of the data $D^{(m)}$ for the current study under model m and $p(m)$ denotes the prior probability of model m in equation (2.8), which is estimated by equation (3.1). The marginal density $p(D^{(m)}|m)$ is precisely the normalizing constant of the joint posterior density of $(\beta^{(m)}, a_0)$.

Computing the posterior model probability $p(m|D^{(m)})$ given in equation (3.3) requires a Monte Carlo method which is different from the method for computing the prior model probability $p(m)$ given in equation (2.8). See Appendix A for an explanation. Let $p(\beta^{(m)}, a_0|D^{(m)})$ denote the joint posterior density of $(\beta^{(m)}, a_0)$, and let $\pi(\beta^{(-m)}|D^{(\mathcal{K})})$ and $p(\beta^{(-m)}|D^{(\mathcal{K})})$ denote the respective marginal prior and posterior densities of $\beta^{(-m)}$ obtained from the full model. Then it can be shown that the posterior probability of model m is given by

$$p(m|D^{(m)}) = \frac{p(\beta^{(-m)}=0|D^{(\mathcal{K})})}{\pi(\beta^{(-m)}=0|D_0^{(\mathcal{K})})}\,p(m) \Big/ \sum_{j=1}^{\mathcal{K}} \frac{p(\beta^{(-j)}=0|D^{(\mathcal{K})})}{\pi(\beta^{(-j)}=0|D_0^{(\mathcal{K})})}\,p(j), \tag{3.5}$$

$m = 1, \ldots, \mathcal{K}$, where $\pi(\beta^{(-m)}=0|D_0^{(\mathcal{K})})$ and $p(\beta^{(-m)}=0|D^{(\mathcal{K})})$ are the marginal prior and posterior densities of $\beta^{(-m)}$ evaluated at $\beta^{(-m)}=0$. In equation (3.5), for notational convenience we let $p(\beta^{(-\mathcal{K})}=0|D^{(\mathcal{K})}, D_0^{(\mathcal{K})}) = 1$. A derivation of result (3.5) is given in Appendix A. In equation (3.5), we use equation (3.1) to compute the prior model probability $p(m)$. Because of the complexity of the prior and posterior distributions, the analytical forms of $\pi(\beta^{(-m)}|D_0^{(\mathcal{K})})$ and $p(\beta^{(-m)}|D^{(\mathcal{K})})$ are not available. However, we can adopt the importance-weighted marginal density estimation (IWMDE) method of Chen (1994) to estimate these marginal prior and posterior densities. The IWMDE method is a Monte Carlo method developed by Chen (1994) which is particularly suitable for estimating marginal posterior densities when the joint posterior density is known up to a normalizing constant. The IWMDE method requires only two respective Markov chain Monte Carlo samples from the prior and posterior distributions for the full model, making the computation of complicated posterior model probabilities feasible. It directly follows from the IWMDE that a simulation consistent estimator of $p(\beta^{(-m)}=0|D^{(\mathcal{K})})$ is given by

$$\hat{p}(\beta^{(-m)}=0|D^{(\mathcal{K})}) = \frac{1}{N}\sum_{l=1}^{N} w(\beta_{(l)}^{(-m)}|\beta_{(l)}^{(m)}, a_{0(l)})\,\frac{p(\beta_{(l)}^{(m)}, \beta^{(-m)}=0, a_{0(l)}|D^{(\mathcal{K})})}{p(\beta_{(l)}^{(\mathcal{K})}, a_{0(l)}|D^{(\mathcal{K})})}, \tag{3.6}$$

where $w(\beta^{(-m)}|\beta^{(m)}, a_0)$ is a completely known conditional density of $\beta^{(-m)}$ given $\beta^{(m)}$ and a_0, and $\{(\beta_{(l)}^{(\mathcal{K})}, a_{0(l)}), l = 1, 2, \ldots, N\}$ is a sample from the joint posterior distribution $p(\beta^{(\mathcal{K})}, a_0|D^{(\mathcal{K})})$ of $(\beta^{(\mathcal{K})}, a_0)$. To construct a good $w(\beta^{(-m)}|\beta^{(m)}, a_0)$, we can use a procedure similar to that used to construct $w(\beta^{(-m)}|\beta^{(m)})$ in equation (3.1) for calculating the prior model probabilities. Similarly, we can obtain $\hat{\pi}(\beta^{(-m)}=0|D_0^{(\mathcal{K})})$, an estimate of $\pi(\beta^{(-m)}=0|D_0^{(\mathcal{K})})$, by using a sample from the joint prior distribution $\pi(\beta^{(\mathcal{K})}, a_0|D_0^{(\mathcal{K})})$.

4. Examples

4.1. Example 1: simulation study

We present a simulation study to demonstrate the methodology. Our main goal here is to demonstrate the behaviour of the prior and posterior model probability structures by using various choices of prior parameters. We show that our method consistently yields the largest posterior probability for the true model under various choices of prior parameters, whereas criterion-based procedures such as the AIC and BIC do not.

We simulate two data sets. The first data set represents the current study and the second data set represents the previous study. For the current study, $n = 200$ independent Bernoulli observations are simulated with success probability

$$p_i = \frac{\exp(-1.0 - 0.5x_{i1} - 2.0x_{i3})}{1 + \exp(-1.0 - 0.5x_{i1} - 2.0x_{i3})}, \qquad i = 1, \ldots, n, \tag{4.1}$$

where x_{i1} and x_{i3} are independent and identically distributed normal random variables with means 1.0 and 0.8 and variances 1.0 and 0.8 respectively. Thus, the true model contains the covariates (x_1, x_3). The average success probability from the 200 generations is 0.121. Two additional covariates (x_{i2}, x_{i4}) are randomly generated, such that the joint distribution of $x_i = (x_{i1}, \ldots, x_{i4})'$ is $N_4(\mu, \Sigma)$, where $\mu = (1.0, 0.5, 0.8, 1.4)$ and

$$\Sigma = \begin{pmatrix} 1.0 & 0.353 & 0 & 0 \\ 0.353 & 0.5 & 0 & 0 \\ 0 & 0 & 0.8 & 0.588 \\ 0 & 0 & 0.588 & 1.2 \end{pmatrix}.$$

Thus the full model consists of the four covariates $(x_1, \ldots, x_4)$ and the true model contains (x_1, x_3). For the previous study, $n_0 = 400$ Bernoulli observations are generated with success probability

$$p_{0i} = \frac{\exp(-1.0 - 1.5x_{0i1} - 0.8x_{0i3})}{1 + \exp(-1.0 - 1.5x_{0i1} - 0.8x_{0i3})}, \qquad i = 1, \ldots, n_0, \tag{4.2}$$

where (x_{0i1}, x_{0i3}) have the same distribution as (x_{i1}, x_{i3}). In addition, two additional covariates, (x_{0i2}, x_{0i4}), are generated such that $x_{0i} = (x_{0i1}, x_{0i2}, x_{0i3}, x_{0i4})'$ has the same distribution as x_i. The average success probability for the previous study is 0.100. An intercept is included in all models, and thus the model space $\mathcal{M}$ consists of 16 models.

The SAS stepwise logistic procedure (SAS Institute, 1989) identifies the (x_3) model as the best model using an entry and exit p-value criterion of 0.2. In addition, the top two models based on the AIC and BIC criteria are given in Table 1. For model m, the AIC and BIC are given by

$$\mathrm{AIC}_m = -2 \log\{L(\hat{\beta}^{(m)}|D^{(m)})\} + 2k_m,$$
$$\mathrm{BIC}_m = -2 \log\{L(\hat{\beta}^{(m)}|D^{(m)})\} + k_m \log(n).$$

Table 1. Top two models based on the AIC and BIC

Model	*AIC*	*BIC*
(x_3)	99.2	105.8
(x_1, x_3)	99.6	109.5

Table 2. Posterior model probabilities for $(\mu_0, \sigma_0^2) = (0.5, 0.008)$, $d_0 = 5$ and various choices of c_0

c_0	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
5	(x_1, x_3)	0.17	0.22	0.31
10	(x_1, x_3)	0.17	0.19	0.35
50	(x_1, x_3)	0.17	0.15	0.38
100	(x_1, x_3)	0.17	0.14	0.39

Table 1 indicates that both the AIC and the BIC identify the same top two models, with the best model being (x_3), and the true model as the second-best model. Thus, both of these criteria fail to identify the true model as the criterion minimizing model.

Now we consider the fully Bayesian approach using prior (2.3). We demonstrate three types of sensitivity analyses in this simulation. We conduct sensitivity analyses with respect to

(a) c_0,
(b) d_0 and
(c) (μ_0, σ_0^2).

These are shown in Tables 2, 3 and 4 respectively. For $\pi_0(\beta^{(m)}|\cdot)$, we use a normal density as given by equation (2.6). To compute the prior and posterior model probabilities, 50000 Gibbs iterations were used to achieve convergence.

Table 2 gives the model with the largest posterior probability using $(\mu_0, \sigma_0^2) = (0.5, 0.008)$, (i.e. $\delta_0 = \lambda_0 = 15$) for several values of c_0. For each value of c_0 in Table 2, the true model, (x_1, x_3), obtains the largest posterior probability, and thus model choice is not sensitive to these values. Moreover, we see that the posterior model probabilities exhibit a monotonic increase as c_0 increases. For example, the posterior model probability for (x_1, x_3) increases 8% from $c_0 = 5$ to $c_0 = 100$. Model (x_3), which is the top model according to the AIC and BIC, has posterior probability less than 10^{-6} for $c_0 \geqslant 5$. Large values of c_0 imply a flat $\pi_0(\beta^{(m)}|c_0)$, for calculating the marginal distribution of the data. Although not shown in Table 2, values of $c_0 < 5$ do not yield (x_1, x_3) as the top model. For example, under the settings of Table 1 with $c_0 = 3$, the top model is (x_1, x_3, x_4) with posterior probability of 0.33, and (x_1, x_3) is the second-best model with posterior probability 0.24. As $c_0 \to 0$ other models obtain the largest posterior probability. For example, when $c_0 = 0.01$, the top model is (x_1, x_2, x_3, x_4) with posterior probability 0.72 and (x_1, x_3) is the fourth-best model with posterior probability 0.01. Thus, model choice becomes sensitive to the choice of c_0 when $c_0 < 5$. We mention that for Table 2, if $c_0 \geqslant 10$, and a uniform prior is used for the model space (i.e. $p(m) = 1/16$), then the true model does not obtain the largest posterior model probability. For example, when $c_0 = 10$ the top model is (x_3) with posterior model probability 0.28. This combination, as mentioned in Section 2.4, yields results that are consistent with those of the AIC and BIC. We see that the wrong model can be obtained when we use a uniform prior on the model space, whereas the true model is obtained when we use the proposed methodology to obtain the prior model probabilities.

From Table 3, we see how the prior model probability is affected as d_0 is changed. In each case, the true model obtains the largest posterior probability. Under the settings of Table 3, the true model also obtains the largest prior probability when $d_0 \geqslant 50$. Table 3 indicates a monotonic increase in the prior (and posterior) model probability as d_0 increases. When $d_0 \geqslant 5$, model choice is not sensitive to the choice of d_0. When $d_0 \geqslant 5$, the (x_3) model has posterior probability less than 10^{-6}. With values of $d_0 < 5$, model choice is sensitive to the

Table 3. Posterior model probabilities for $(\mu_0, \sigma_0^2) = (0.5, 0.008)$, $c_0 = 5$ and various choices of d_0

d_0	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
10	(x_1, x_3)	0.29	0.22	0.45
50	(x_1, x_3)	0.54	0.22	0.69
100	(x_1, x_3)	0.59	0.22	0.76

Table 4. Posterior model probabilities for $c_0 = 10$, $d_0 = 5$ and various choices of (μ_0, σ_0^2)

(μ_0, σ_0^2)	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
(0.5, 0.083)	(x_1, x_3)	0.17	0.23	0.31
(0.5, 0.023)	(x_1, x_3)	0.17	0.24	0.34
(0.5, 0.008)	(x_1, x_3)	0.17	0.19	0.35
$(0.98, 3.7 \times 10^{-4})$	(x_1, x_3)	0.17	0.08	0.36

choice of d_0. For example, when $d_0 = 3$, the top model is (x_1, x_2, x_3, x_4) with prior probability 0.44 and posterior probability 0.33. The true model, (x_1, x_3), is the fourth-best model with prior probability 0.08 and posterior probability 0.17. As $d_0 \to 0$, the prior on the model space approaches a uniform prior, and in this case we obtain top models similar to those obtained by the AIC and BIC. For example, when $d_0 = 10^{-4}$, $c_0 = 10$ and $(\mu_0, \sigma_0^2) = (0.5, 0.008)$, the top model is (x_3) with prior model probability 0.058 and posterior model probability 0.25. The true model, (x_1, x_3), is the second-best model with prior model probability 0.065 and posterior model probability 0.19. As both c_0 and d_0 become large, the true model obtains the top prior and posterior model probability. For example, with $c_0 = d_0 = 100$, the prior model probability for the true model is 0.59 and the posterior probability is 0.81. Thus, we see how sharp the prior and posterior model probabilities become as c_0 and d_0 are increased.

Table 4 shows a sensitivity analysis with respect to (μ_0, σ_0^2). The true model obtains the largest posterior probability in each case. Under these settings, model choice is not sensitive to the choice of (μ_0, σ_0^2). There is a monotonic increase in the posterior model probability as more weight is given to the historical data. Moreover, the true model obtains the largest posterior probability even when a uniform prior (i.e. $(\mu_0, \sigma_0^2) = (0.5, 0.083)$) is specified for a_0. In addition, when c_0 and d_0 are increased, the posterior probability of the true model is increased. For example, with $c_0 = d_0 = 10$ and a uniform prior on a_0, the posterior probability of (x_1, x_3) is 0.45. Under the settings of Table 4, model (x_3) has posterior probability less than 10^{-6}.

In summary, Tables 2–4 show that, if c_0 and d_0 are moderately large, then model choice is not sensitive to their choices. However, model choice can be sensitive if c_0 and/or d_0 are taken to be small. Moreover, we see monotonic increases in the posterior model probability as c_0 or d_0 are increased. We also observe that model choice is not as sensitive to the choice of (μ_0, σ_0^2) (Table 4). The true model obtains the highest posterior probability for a wide range of (μ_0, σ_0^2), and the posterior model probability increases as more weight is given to the historical data. To obtain results that are consistent with the AIC and BIC, we let $c_0 \to \infty$ and $d_0 \to 0$, and we let $a_0 \to 0$ with probability 1. In the simulation above, we obtained results consistent with the AIC and BIC by letting $c_0 \to \infty$ and $d_0 \to 0$, and taking moderately non-informative choices for (μ_0, σ_0^2). Finally, we remind the reader that the AIC and BIC computations are based only on the current data and do not use the historical data, as in the

Table 5. Summary of the ACTG019 trial data

	x_{01} *(count)*	x_{02} *(years)*	x_{03} *(frequency)*		x_{04} *(frequency)*		y_0 *(frequency)*	
Mean	334.7	34.64	AZT	418	White	752	1	55
Standard deviation	109.3	7.679	Placebo	405	Other	71	0	768

Table 6. Summary of the ACTG036 trial data

	x_1 *(count)*	x_2 *(years)*	x_3 *(frequency)*		x_4 *(frequency)*		x_5 *(frequency)*		x_6 *(frequency)*		y *(frequency)*	
Mean	297.7	30.43	AZT	89	White	166	VIII	163	Yes	78	1	11
Standard deviation	130.5	11.16	Placebo	94	Other	17	Other	20	No	105	0	172

proposed variable selection method. In general, it is possible that the AIC and BIC could obtain the correct model as the top model if we do an analysis based on combining the historical data and the current data in one data set.

4.2. Example 2: study of acquired immune deficiency syndrome

We consider an analysis of the AIDS study ACTG036 using the data from the ACTG019 study as prior information. The purpose of this example is to demonstrate the methodology proposed and to show that, by incorporating prior information from ACTG019, results can be obtained that are different from those of criterion-based procedures such as the AIC and BIC.

The ACTG019 study was a double-blind placebo-controlled clinical trial comparing zidovudine (AZT) with a placebo in people with CD4 cell 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 ARC and 0 otherwise. Several covariates were also measured. Those which we use here are the CD4 cell count (x_{01}) (cell count per cubic millimetre of serum), age (x_{02}), treatment (x_{03}) and race (x_{04}). The covariates CD4 cell count and age are continuous, whereas the other covariates are binary. Table 5 summarizes the covariates and the response variable for the ACTG019 study. The ACTG036 study was also a placebo-controlled clinical trial comparing AZT with a placebo in patients with hereditary coagulation disorders. The results of this study have been published by Merigen *et al.* (1991). The sample size in this study, excluding cases with missing data, was $n = 183$. The response variable (y) for these data is binary with 1 indicating death, development of AIDS or ARC and 0 otherwise. Several covariates were measured for these data. Those which we use here are the CD4 cell count (x_1), age (x_2), treatment (x_3), race (x_4), haemophilia factor type (x_5) and monoclonal factor concentrate use (x_6). The covariates CD4 cell count and age are continuous, whereas the other covariates are binary. The covariate x_5 was coded 1 if the haemophilia factor type was factor VIII and 0 otherwise, and the covariate x_6 was coded 1 if the patient used a monoclonal factor concentrate and 0 otherwise. Table 6 summarizes the covariates and the response variable for the ACTG036 study.

Table 7. Top two models based on the AIC and BIC

Model	*AIC*	*BIC*
(x_1)	65.8	72.3
(x_1, x_2)	67.6	77.2

Table 8. Posterior model probabilities for $d_0 = 0.5$, $(\mu_{01}, \sigma_{01}^2) = (0.5, 0.005)$ and $(\mu_{02}, \sigma_{02}^2) = (0.5, 0.083)$ for various choices of c_0

c_0	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
3	(x_1, x_2, x_3)	0.057	0.053	0.120
5	(x_1, x_2, x_3)	0.057	0.059	0.121
10	(x_1, x_2, x_3)	0.057	0.061	0.122

The SAS stepwise logistic procedure identifies the (x_1) model as the best model for the ACTG036 data using an entry and exit p-value criterion of 0.2. Table 7 gives the top two models based on the AIC and BIC, which also identify (x_1) as the best model. Since the covariates in the previous study (ACTG019) are a subset of the covariates in the current study (ACTG036), we use the priors developed in expression (2.5) to carry out the selection of the subset of variables. An intercept is included in every model and 50000 Gibbs iterations were used to achieve convergence.

Table 8 shows the results of a Bayesian selection of variables for several values of c_0 using $d_0 = 0.5$, $(\mu_{01}, \sigma_{01}^2) = (0.5, 0.005)$ (i.e. $\delta_{01} = \lambda_{01} = 25$) and $(\mu_{02}, \sigma_{02}^2) = (0.5, 0.083)$ (i.e. $\delta_{02} = \lambda_{02} = 1$). We see that for this choice the top model is (x_1, x_2, x_3), which is quite different from the model selected by the AIC and BIC. The second-best model is (x_1, x_2) with a posterior model probability of 0.08. Thus, we see that with a moderate incorporation of the historical data and a nearly uniform prior on the model space (i.e. $d_0 = 0.5$) a top model is obtained which is different from that given by the AIC and BIC. We note that for $c_0 = 3$ the (x_1) model has a posterior model probability of 0.024 and the (x_1, x_2) model has a posterior probability of 0.08. For $c_0 \geqslant 3$ and $d_0 \geqslant 0.5$, the top model remains (x_1, x_2, x_3) and its prior and posterior probabilities increase monotonically as (d_0, c_0) increase. For example, when $c_0 = 3$ and $d_0 = 5$, the prior probability of model (x_1, x_2, x_3) is 0.17 and its posterior probability is 0.28. When $c_0 = 50$ and $d_0 = 10$, the prior model probability is 0.20 and the posterior probability is 0.31. When $d_0 \geqslant 5$, the (x_1, x_2, x_3) model also obtains the largest prior model probability. We see that model choice is not sensitive when $c_0 \geqslant 3$ and $d_0 \geqslant 0.5$, for which (x_1, x_2, x_3) is consistently the top model. To obtain a result that is similar to that from the AIC and BIC, we let $d_0 \to 0$. With $d_0 = 0.001$, and $c_0 = 50$, the top model is (x_1) with posterior model probability 0.22. In general, we see that our prior elicitation scheme is quite flexible and can obtain criterion-based results with appropriate choices of prior parameters.

Table 9 shows the results of a sensitivity analysis on a_{02}. In each case, the (x_1, x_2, x_3) model obtains the largest posterior probability and the results are not sensitive to the choice of prior parameters. We observe a monotonic decrease in the posterior probability of model (x_1, x_2, x_3) as more prior weight is given to the part of the prior that incorporates the new covariates. This is expected since the new covariates in the current study do not involve age (x_2) or treatment (x_3). Tables 8 and 9 demonstrate that the incorporation of the ACTG019 trial data into the current analysis reveals the importance of age (x_2) and treatment (x_3) as well as CD4 cell count (x_1) in predicting the response. By incorporating the prior information from the

Table 9. Posterior model probabilities for $d_0 = 3$, $c_0 = 10$, $(\mu_{01}, \sigma_{01}^2) = (0.5, 0.005)$ and various choices of $(\mu_{02}, \sigma_{02}^2)$

$(\mu_{02}, \sigma_{02}^2)$	m	$p(m)$	$p(D\|m)$	$p(m\|D)$
(0.5, 0.083)	(x_1, x_2, x_3)	0.14	0.06	0.24
(0.5, 0.023)	(x_1, x_2, x_3)	0.14	0.06	0.23
$(0.98, 3.7 \times 10^{-4})$	(x_1, x_2, x_3)	0.14	0.05	0.20

ACTG019 study, the model with the largest posterior probability includes two covariates, treatment and age, which are not included in the best model based on the AIC and BIC procedures. Moreover, after incorporation of the prior information from ACTG019, model (x_1) has very small posterior probabilities for several choices of c_0 and d_0. With an analysis of the ACTG019 trial data alone, the BIC and the stepwise procedure yield the model containing age, treatment and CD4 cell count as the best model. Thus the prior distributions incorporate the importance of these two additional covariates (age and treatment) into the ACTG036 analysis, and as a result have an effect on the final results. The result obtained here is important since it shows the importance of the treatment and age covariates when the prior information from a previous study is incorporated. Such results play a crucial role in decision-making and public policy regarding the treatment of AIDS. Such a result would not have been obtained by using criterion-based methods such as the AIC and BIC. In fact, for the ACTG036 trial data, the (x_1, x_2, x_3) model obtained the ninth-smallest AIC and BIC values. We see in this analysis how the incorporation of data from a previous study can affect the model choice and yield results that are different from those of an analysis that uses criterion-based methods.

5. Discussion

We have developed a general class of prior distributions for the logistic regression model and have also derived some novel computational tools. The main focus of application for our methods has been the selection of variables, but the priors, as well as the computational methods, can be used for other applications.

For the AIDS data discussed in Section 4, we saw how the incorporation of the prior information affected the results of the current study. This is important, since it may shed light on the importance of certain prognostic factors that would not otherwise have been identified in the analysis. Such information is especially crucial for an epidemic like AIDS, since it may facilitate policy and decision-making regarding the treatment of the disease.

We mention that the methods of George *et al.* (1996) and Raftery (1996) are quite different from the methods proposed here. George *et al.* (1996) took the priors for the regression coefficients to be normal scale mixtures. Their priors are well suited for very large problems and are mainly designed for tuning the convergence of the Gibbs sampler, whereas the priors proposed here attempt to incorporate real prior information for model selection. George *et al.* (1996) did not address informative prior elicitation or the quantification of real prior information, whereas our focus has been more on the synthesis of informative priors and the quantification of prior information from past studies. With our proposed priors, the computational methods advanced here are well suited to handle a moderate number of covariates, such as $k \leqslant 20$. The methods of Raftery (1996) are based on asymptotic approximations to the Bayes factor for generalized linear models. Raftery (1996) did not discuss prior elicitation or Markov chain Monte Carlo methods to compute Bayes factors or posterior model

probabilities. We have avoided asymptotic approximations of any sort and do all the computations by using novel Markov chain Monte Carlo techniques. Our methods can be used in many practical situations. For example 1 ($k = 4$), the computing time required to obtain posterior probabilities for all possible models was about 40 min using a Digital alpha machine with a single processor.

Our priors (2.3), (2.5) and (2.8) overcome the sensitivity issue to a certain extent by their construction. As demonstrated in the numerical examples of Section 4, our priors are quite robust under a wide variety of prior parameters. A nice property of our prior (2.3) is that as $c_0 \to \infty$ the prior remains proper, and thus the prior cannot become vague by altering these parameters. This is a solid feature demonstrated in examples 1 and 2 of Section 4. Although our priors are fairly robust, model choice can be sensitive to certain choices of prior parameters as demonstrated in Section 4. The sensitivity, however, is *not* because prior (2.3) approaches impropriety, since it is always proper as long as we have a proper non-degenerate prior on a_0 (see Chen *et al.* (1997)). If $a_0 \to 0$ *and* $c_0 \to \infty$ then prior (2.3) tends to an improper prior, but such a scenario does not make any practical sense. Hence, our priors are better protected against impropriety and are more stable over certain ranges of prior parameters, but they are nevertheless not immune to sensitivity. Sensitivity and robustness are an important issue which is relevant in our proposed methodology.

Acknowledgements

The authors wish to thank the Joint Editor, the Associate Editor and three referees for several suggestions which have greatly improved the paper. Dr Chen's research was supported by National Science Foundation grant DMS-9702172, and Dr Ibrahim's research was supported by National Institutes of Health grants CA 70101-01 and CA 74015-01.

Appendix A: Computational development

We first describe how to sample from the joint posterior distribution of $(\beta^{(\mathcal{K})}, a_0|D^{(\mathcal{K})})$ for the full model. Given a_0, the posterior density of $\beta^{(\mathcal{K})}$ is log-concave in each component. Thus random variate generation from these univariate distributions is readily accomplished by using the adaptive rejection algorithm of Gilks and Wild (1992) for the Gibbs sampler. The conditional posterior density of a_0 is of the form

$$p(a_0|\beta^{(\mathcal{K})}, D^{(\mathcal{K})}) \propto \exp\{a_0(y_0' X_0 \beta^{(\mathcal{K})} - J_0' Q_0)\} a_0^{\delta_0 - 1}(1 - a_0)^{\lambda_0 - 1}. \tag{A.1}$$

Since $p(a_0|\beta^{(\mathcal{K})}, D^{(\mathcal{K})})$ is not log-concave in general, we propose the following Metropolis algorithm to sample a_0. Consider

$$a_0 = \frac{\exp(\xi)}{1 + \exp(\xi)}. \tag{A.2}$$

Then, the conditional posterior distribution $(\xi|\beta^{(\mathcal{K})}, D^{(\mathcal{K})})$ is

$$p(\xi|\beta^{(\mathcal{K})}, D^{(\mathcal{K})}) \propto p(a_0|\beta^{(\mathcal{K})}, D^{(\mathcal{K})}) \frac{\exp(\xi)}{\{1 + \exp(\xi)\}^2}, \tag{A.3}$$

where $p(a_0|\beta^{(\mathcal{K})}, D^{(\mathcal{K})})$ is given by expression (A.1) and a_0 is evaluated at $a_0 = \exp(\xi)/\{1 + \exp(\xi)\}$. Instead of directly generating a_0 from expression (A.1), we first generate ξ from expression (A.3) and then use equation (A.2) to obtain a_0. To generate ξ, we use a normal proposal $N(\hat{\xi}, \hat{\tau}_{\xi}^2)$, where $\hat{\xi}$ is a maximizer of the logarithm of the right-hand side of expression (A.3). Also, $\hat{\tau}_{\xi}^2$ is minus the inverse of the second derivative of $\log\{p(\xi|\beta^{(\mathcal{K})}, D^{(\mathcal{K})})\}$ evaluated at $\xi = \hat{\xi}$, given by

$$\hat{\tau}_{\xi}^{-2} = -\left.\frac{\mathrm{d}^2[\log\{p(\xi|\beta^{(K)}, D^{(K)})\}]}{\mathrm{d}\xi^2}\right|_{\xi=\hat{\xi}}.$$

The algorithm to generate ξ operates as follows. Let ξ be the current value and then generate a proposal value ξ^* from $N(\hat{\xi}, \hat{\tau}_{\xi}^2)$. A move from ξ to ξ^* is made with probability

$$\min\left\{\frac{p(\xi^*|\beta^{(K)}, D^{(K)})\,\phi\{(\xi-\hat{\xi})/\hat{\tau}_{\xi}\}}{p(\xi|\beta^{(K)}, D^{(K)})\,\phi\{(\xi^*-\hat{\xi})/\hat{\tau}_{\xi}\}}, 1\right\},$$

where ϕ is the standard normal probability density function. After we have obtained ξ, we compute a_0 by using equation (A.2). This procedure yields samples from the joint posterior distribution of $(\beta^{(K)}, a_0)$. Since the joint prior distribution $\pi(\beta^{(K)}, a_0|D_0^{(K)})$ is of the same exact form as the joint posterior, the same algorithm can be used to generate samples from the joint prior distribution of $(\beta^{(K)}, a_0)$. Finally, we note that generating $\beta^{(K)}$ from $p_0(\beta^{(K)}|D_0^{(K)})$ given in expression (3.2) is even simpler since this random generation requires only the adaptive rejection algorithm of Gilks and Wild (1992).

The technical details for computing the prior model probabilities $\hat{p}(m)$, given in equation (3.1), are as follows. Suppose that under the full model we have a sample $\{\beta_{0(l)}^{(K)}, l = 1, \ldots, N_0\}$ from $p_0(\beta^{(K)}|D_0^{(K)})$. Let

$$p(D_0^{(m)}|m) = \int p_0^*(\beta^{(m)}|D_0^{(m)})\,\mathrm{d}\beta^{(m)}, \tag{A.4}$$

where $p_0^*(\beta^{(m)}|D_0^{(m)})$ is given in equation (2.7). Then, using the result of Chen and Shao (1997a), we have the key identity

$$\frac{p(D_0^{(m)}|m)}{p(D_0^{(K)}|K)} = E\left\{\frac{p_0^*(\beta^{(m)}|D_0^{(m)})\,w(\beta^{(-m)}|\beta^{(m)})}{p_0^*(\beta^{(K)}|D_0^{(K)})}\right\}, \tag{A.5}$$

where the expectation is taken with respect to the density $p_0(\beta^{(K)}|D_0^{(K)})$. Note that the choice of the weight function $w(\beta^{(-m)}|\beta^{(m)})$ is somewhat arbitrary. However, Chen and Shao (1997a) showed that the best choice of $w(\beta^{(-m)}|\beta^{(m)})$ is the conditional density of $\beta^{(-m)}$ given $\beta^{(m)}$ with respect to the density $p_0(\beta^{(K)}|D_0^{(K)})$. Since the closed form expression of this conditional density is not available, we follow an empirical procedure provided by Chen (1994) to select $w(\beta^{(-m)}|\beta^{(m)})$. Specifically, using the sample $\{\beta_{0(l)}^{(K)}, l = 1, \ldots, N_0\}$, we construct the mean and covariance matrix, denoted by $(\tilde{\beta}_0, \tilde{\Sigma}_0)$, and then we choose $w(\beta^{(-m)}|\beta^{(m)})$ to be the conditional density of the k-dimensional normal distribution, $N_k(\tilde{\beta}_0, \tilde{\Sigma}_0)$, for $\beta^{(-m)}$ given $\beta^{(m)}$. A nice feature of this procedure is that $w(\beta^{(-m)}|\beta^{(m)})$ is calculated automatically. See Ibrahim *et al.* (1996) for more details regarding $w(\beta^{(-m)}|\beta^{(m)})$.

It can be seen that, using equation (A.5), $p(m)$ can be rewritten as

$$p(m) = E\left\{\frac{p_0^*(\beta^{(m)}|D_0^{(m)})\,w(\beta^{(-m)}|\beta^{(m)})}{p_0^*(\beta^{(K)}|D_0^{(K)})}\right\}\Big/\sum_{j=1}^{K} E\left\{\frac{p_0^*(\beta^{(j)}|D_0^{(j)})\,w(\beta^{(-j)}|\beta^{(j)})}{p_0^*(\beta^{(K)}|D_0^{(K)})}\right\}. \tag{A.6}$$

Thus, the estimate of $p(m)$ given in equation (3.1) follows directly from equation (A.6).

Now, we consider calculating posterior model probabilities. We first explain the reason why we require a Monte Carlo method for computing the posterior model probability $p(m|D^{(m)})$ given in equation (3.3) that is different from the method for computing the prior model probability $p(m)$ given in equation (2.8). From equation (3.4), it can be seen that the calculation of posterior probabilities requires evaluating

$$\begin{aligned} p(D^{(m)}|m) &= \int L(\beta^{(m)}|D^{(m)})\,\pi(\beta^{(m)}, a_0|D_0^{(m)})\,\mathrm{d}\beta^{(m)}\,\mathrm{d}a_0 \\ &= \int L(\beta^{(m)}|D^{(m)})\,\frac{\pi^*(\beta^{(m)}, a_0|D_0^{(m)})}{c_m}\,\mathrm{d}\beta^{(m)}\,\mathrm{d}a_0, \end{aligned} \tag{A.7}$$

where the unnormalized joint prior density

$$\pi^*(\beta^{(m)}, a_0|D_0^{(m)}) = \exp\{a_0(y_0'X_0^{(m)}\beta^{(m)} - J_0'Q_0^{(m)})\}\,\pi_0(\beta^{(m)}|c_0)a_0^{\delta_0-1}(1-a_0)^{\lambda_0-1}, \tag{A.8}$$

and the normalizing constant for the joint prior density is given by

$$c_m = \int \pi^*(\beta^{(m)}, a_0|D_0^{(m)})\,\mathrm{d}\beta^{(m)}\,\mathrm{d}a_0. \tag{A.9}$$

Owing to the complexity of equation (A.8), the closed form of c_m does not appear possible. Therefore, computing $p(D^{(m)}|m)$ requires evaluating the ratio of two analytically intractable integrals, which is essentially a ratio of two normalizing constants. See, for example, Chen and Shao (1997b). However, to compute each of similar quantities involved in the prior model probability $p(m)$ given in equation (2.8), we only need to evaluate one integral, i.e.

$$\int \exp(y_0'X_0^{(m)}\beta^{(m)} - J_0'Q_0^{(m)})\,\pi_0(\beta^{(m)}|d_0)\,\mathrm{d}\beta^{(m)},$$

because a closed form of $\pi_0(\beta^{(m)}|d_0)$ is available.

To derive equation (3.5) we need several key theoretical results.

Lemma 1. Under model m, Let

$$\pi(\beta^{(m)}|D_0^{(m)}) = \int \pi(\beta^{(m)}, a_0|D_0^{(m)})\,\mathrm{d}a_0$$

denote the marginal prior distribution of $\beta^{(m)}$ and let

$$p(\beta^{(m)}|D^{(m)}) = \int p(\beta^{(m)}, a_0|D^{(m)})\,\mathrm{d}a_0$$

denote the marginal posterior distribution of $\beta^{(m)}$. Then, for any model $m \in \mathcal{M}$,

$$p(D^{(m)}|m) = \frac{L(\beta^{(m)}|D^{(m)})\,\pi(\beta^{(m)}|D_0^{(m)})}{p(\beta^{(m)}|D^{(m)})}, \tag{A.10}$$

for all $\beta^{(m)}$.

Lemma 2. From expressions (2.1) and (2.3) and the construction of $\pi_0(\beta^{(m)}|c_0)$, we have

(a) $L(\beta^{(m)}|D^{(m)}) = L(\beta^{(m)}, \beta^{(-m)} = 0|D^{(\mathcal{K})})$ where $L(\beta^{(m)}, \beta^{(-m)} = 0|D^{(\mathcal{K})})$ is the likelihood function for the full model evaluated at $\beta^{(\mathcal{K})} = (\beta^{(m)}, \beta^{(-m)} = 0)$,
(b) $p(\beta^{(m)}|D^{(m)}) = p(\beta^{(m)}|\beta^{(-m)} = 0, D^{(\mathcal{K})})$ where $p(\beta^{(m)}|\beta^{(-m)} = 0, D^{(\mathcal{K})})$ is the conditional posterior density of $\beta^{(m)}$ given $\beta^{(-m)} = 0$ obtained from the marginal posterior density based on the full model, $p(\beta^{(\mathcal{K})}|D^{(\mathcal{K})})$, and
(c) $\pi(\beta^{(m)}|D_0^{(m)}) = \pi(\beta^{(m)}|\beta^{(-m)} = 0, D_0^{(\mathcal{K})})$ where $\pi(\beta^{(m)}|\beta^{(-m)} = 0, D_0^{(\mathcal{K})})$ is the conditional prior distribution of $\beta^{(m)}$ given $\beta^{(-m)} = 0$ obtained from the marginal prior density based on the full model, $\pi(\beta^{(\mathcal{K})}|D_0^{(\mathcal{K})})$.

Lemmas 1 and 2 yield the following key identities.

Theorem 1. Let $\pi(\beta^{(-m)}|D_0^{(\mathcal{K})})$ and $p(\beta^{(-m)}|D^{(\mathcal{K})})$ denote the respective marginal prior and posterior distributions of $\beta^{(-m)}$ obtained from the full model. Then

$$\frac{p(D^{(m)}|m)}{p(D^{(\mathcal{K})}|\mathcal{K})} = \frac{p(\beta^{(-m)} = 0|D^{(\mathcal{K})})}{\pi(\beta^{(-m)} = 0|D_0^{(\mathcal{K})})}, \qquad m = 1, \ldots, \mathcal{K}. \tag{A.11}$$

Proof. For the full model (A.10), specifying $\beta^{(\mathcal{K})} = (\beta^{(m)\prime}, \beta^{(-m)\prime} = 0)'$ yields

$$\frac{p(D^{(m)}|m)}{p(D^{(\mathcal{K})}|\mathcal{K})} = \frac{L(\beta^{(m)}|D^{(m)})\,\pi(\beta^{(m)}|D_0^{(m)})/p(\beta^{(m)}|D^{(m)})}{L(\beta^{(m)}, \beta^{(-m)} = 0|D^{(\mathcal{K})})\,\pi(\beta^{(m)}, \beta^{(-m)} = 0|D_0^{(\mathcal{K})})/p(\beta^{(m)}, \beta^{(-m)} = 0|D^{(\mathcal{K})})}. \tag{A.12}$$

Since

$$p(\beta^{(m)}, \beta^{(-m)} = 0|D^{(\mathcal{K})}) = p(\beta^{(-m)} = 0|D^{(\mathcal{K})})\, p(\beta^{(m)}|\beta^{(-m)} = 0, D^{(\mathcal{K})})$$

and

$$\pi(\beta^{(m)}, \beta^{(-m)} = 0|D_0^{(\mathcal{K})}) = \pi(\beta^{(-m)} = 0|D_0^{(\mathcal{K})})\, \pi(\beta^{(m)}|\beta^{(-m)} = 0, D_0^{(\mathcal{K})}),$$

equation (A.11) follows from lemma 2 and equation (A.12), whereas equation (3.5) simply follows from equation (A.11). This completes the proof.

The results given in theorem 1 are very attractive since they show that the posterior probability $p(m|D^{(m)})$ is simply a function of the prior model probabilities $p(m)$ and the marginal prior and posterior density functions of $\beta^{(-m)}$ for the full model evaluated at $\beta^{(-m)} = 0$. Therefore, to estimate the posterior model probabilities for all $m \in \mathcal{M}$, we use equation (3.1) to estimate the prior model probabilities and we compute the marginal prior and posterior density functions only for the full model via the IWMDE method of Chen (1994). A typical formulation of an IWMDE is given in equation (3.6). The performance of an IWMDE depends on the choice of the weight density function $w(\beta^{(-m)}|\beta^{(m)}, a_0)$ given in equation (3.6). A theoretical exploration of a performance study of the IWMDE method can be found in Chen and Shao (1997c). Chen and Shao (1997c) also found that IWMDE is better than a kernel density estimation (for example, see Silverman (1986)) under the Kullback–Leibler divergence. Following the guideline given in Chen (1994), we choose a k-dimensional normal distribution, whose mean and covariance matrix are estimated by the Markov chain Monte Carlo outputs from the joint posterior (or prior) distribution, to construct w. This choice is intuitively appealing since the conditional prior distribution $\pi(\beta^{(\mathcal{K})}|D_0^{(\mathcal{K})}, a_0)$ as well as the conditional posterior distribution $p(\beta^{(\mathcal{K})}|D^{(\mathcal{K})}, a_0)$ are asymptotically normal under certain regularity conditions (for example, see Bernardo and Smith (1994) and Schervish (1995)). However, a special consideration regarding the choice of w should be taken when the sample size is extremely small or the dimension of $\beta^{(\mathcal{K})}$ (i.e. k) is large.

References

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. Int. Symp. Inference Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest: Akademiai Kiàdo.

Albert, J. H. (1988) Computational methods using a Bayesian hierarchical generalized linear model. *J. Am. Statist. Ass.*, **83**, 1037–1044.

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Bedrick, E. J., Christensen, R. and Johnson, W. (1996) A new perspective on priors for generalized linear models. *J. Am. Statist. Ass.*, **91**, 1450–1460.

Berger, J. O. and Pericchi, L. R. (1996) The intrinsic Bayes factor for model selection and prediction. *J. Am. Statist. Ass.*, **91**, 109–122.

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

Chen, M.-H. (1994) Importance-weighted marginal Bayesian posterior density estimation. *J. Am. Statist. Ass.*, **89**, 818–824.

Chen, M.-H., Ibrahim, J. G. and Yiannoutsos, C. (1997) Prior elicitation, variable selection, and Bayesian computation for logistic regression models. *Technical Report*. Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester. (Available from `http://www.wpi.edu/~mhchen/logit`.)

Chen, M.-H. and Shao, Q.-M. (1997a) Estimating ratios of normalizing constants for densities with different dimensions. *Statist. Sin.*, **7**, 607–630.

———(1997b) On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.*, **25**, 1563–1594.

———(1997c) Performance study of marginal posterior density estimation via Kullback-Leibler divergence. *Test*, **6**, 321–350.

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

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1996) Efficient parametrisations for generalized linear mixed models (with discussion). In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 165–180. Oxford: Oxford University Press.

George, I. E. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

George, I. E., McCulloch, R. E. and Tsay, R. S. (1996) Two approaches to Bayesian model selections with applications. In *Bayesian Analysis in Econometrics and Statistics—Essays in Honor of Arnold Zellner* (eds D. A. Berry, K. A. Chaloner and J. K. Geweke), pp. 339–348. New York: Wiley.

Gilks, W. R. and Wild, P. (1992) Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Ibrahim, J. G., Chen, M.-H. and MacEachern, S. N. (1996) Bayesian variable selection for proportional hazards models. *Technical Report*. Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston.

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

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

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-ACTG 036 Study Group (1991) Placebo-controlled trial to evaluate zidovudine in treatment of human immunodeficiency virus infection in asymptomatic patients with hemophilia. *Blood*, 900–906.

Müller, P. and Roeder, K. (1996) A Bayesian semiparametric model for case-control studies with errors in variables. *Personal Communication*.

Raftery, A. E. (1996) Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, **83**, 251–266.

SAS Institute (1989) *SAS/STAT User's Guide, Version 6*, 4th edn. Cary: SAS Institute.

Schervish, M. (1995) *Theory of Statistics*. New York: Springer.

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

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

Smith, A. F. M. and Spiegelhalter, D. J. (1980) Bayes factors and choice criteria for linear models. *J. R. Statist. Soc.* B, **42**, 213–220.

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., Hyslop, 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 Engl. J. Med.*, **322**, 941–949.

West, M. (1985) Generalized linear models: scale parameters, outlier accommodations and prior distributions. In *Bayesian Statistics 2* (eds J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith). Amsterdam: North-Holland.

West, M., Harrison, P. J. and Migon, H. S. (1985) Dynamic generalized linear models and Bayesian forecasting (with discussion). *J. Am. Statist. Ass.*, **80**, 73–97.

Zeger, S. L. and Karim, M. R. (1991) Generalized linear models with random effects; a Gibbs sampling approach. *J. Am. Statist. Ass.*, **86**, 79–86.

Zellner, A. and Rossi, P. E. (1984) Bayesian analysis of dichotomous quantal response models. *J. Econometr.*, **25**, 365–393.