Published in final edited form as:

J Am Stat Assoc. 2008 December 1; 103(484): 1659–1664. doi:10.1198/016214508000000779.

Properties and Implementation of Jeffreys's Prior in Binomial Regression Models

Ming-Hui Chen, Joseph G. Ibrahim, and Sungduk Kim

Ming-Hui Chen is Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269 (E-mail: mhchen@stat.uconn.edu). Joseph G. Ibrahim is Alumni Distinguished Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 (E-mail: ibrahim@bios.unc.edu). Sungduk Kim is Research Fellow, Division of Epidemiology, Statistics and Prevention Research, National Institute of Child Health and Human Development, NIH Rockville, MD 20852 (E-mail: kims2@mail.nih.gov).

Abstract

We study several theoretical properties of Jeffreys's prior for binomial regression models. We show that Jeffreys's prior is symmetric and unimodal for a class of binomial regression models. We characterize the tail behavior of Jeffreys's prior by comparing it with the multivariate t and normal distributions under the commonly used logistic, probit, and complementary log–log regression models. We also show that the prior and posterior normalizing constants under Jeffreys's prior are linear transformation-invariant in the covariates. We further establish an interesting theoretical connection between the Bayes information criterion and the induced dimension penalty term using Jeffreys's prior for binomial regression models with general links in variable selection problems. Moreover, we develop an importance sampling algorithm for carrying out prior and posterior computations under Jeffreys's prior. We analyze a real data set to illustrate the proposed methodology.

Keywords

Bayes information criterion; Importance sampling; Normalizing constant; Tail behavior

1. INTRODUCTION

Jeffreys's prior is perhaps the most widely used noninformative prior in Bayesian analysis. For the binomial regression model, Jeffreys's prior is attractive because it is proper under mild conditions and requires no elicitation of hyperparameters whatsoever. The only requirement is a likelihood function from which the prior is then derived using Jeffrey's rule, which is to take the prior distribution to be the determinant of the square root of the Fisher information matrix. There has been an enormous literature on Jeffrey's prior and its properties for a wide variety of applications and models, as well as its connections to various reference priors proposed in the literature. This literature is too vast to list in its entirety here, but some relevant key references include works of Jeffreys (1946, 1961), Kass (1989, 1990), Ibrahim and Laud (1991), Kass and Wasserman (1996), and Berger (2000, 2006). Two excellent books discussing Jeffreys's prior include those of Box and Tiao (1973) and Berger (1985).

Although the literature on Jeffreys's prior is vast, there is very little discussion of theoretical properties of Jeffreys's prior for models in which it is proper, particularly for logistic regression. Such properties include its potential connections to normal or t distributions, the tail behavior of Jeffreys's prior, unimodality and symmetry properties, techniques for sampling, and its properties in variable selection problems. In Section 2 we carry out a detailed investigation of these theoretical properties of Jeffreys's prior for general binomial regression models. In

Section 3 we propose an efficient importance sampling algorithm for computing the prior and posterior normalizing constants for Jeffreys's prior and examine its performance. The development of this importance sampling algorithm completely alleviates the need for more complex, time-consuming, and computationally intensive algorithms, such as Gibbs sampling or other Markov chain Monte Carlo (MCMC) algorithms. In Section 4 we use Jeffreys's prior in analyzing a prostate cancer data set. We give a brief discussion of our results and future research in Section 5. We give all proofs and the computational development in Appendixes A and B of the supplementary document.

2. PROPERTIES OF JEFFREYS'S PRIOR IN BINOMIAL REGRESSION MODELS

Suppose that $\{(\mathbf{x}_i, y_i, n_i), i = 1, 2, \dots, n\}$ are independent observations, where y_i is a binomial response variable taking a value between 0 and n_i (≥ 1), and $\mathbf{x}_i = (x_{i0}, x_{i1}, \dots, x_{ik})'$ is a $(k+1) \times 1$ vector of (possibly random) covariates and $x_{i0} = 1$ for the intercept term. The binomial regression model is assumed for $[y_i|\mathbf{x}_i]$, which has conditional density given by

$$p(y_i|\mathbf{x}_i, n_i, \boldsymbol{\beta}) = \binom{n_i}{y_i} \left\{ F(\mathbf{x}_i' \boldsymbol{\beta}) \right\}^{y_i} \left\{ 1 - F(\mathbf{x}_i' \boldsymbol{\beta}) \right\}^{n_i - y_i}, \quad i=1, 2, \dots, n, \quad (1)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$ denotes a $(k+1)$ vector of regression coefficients, $F(\cdot)$ denotes a cumulative distribution function (cdf), and F^{-1} is called the link function. The likelihood function of $\boldsymbol{\beta}$ is

$$L(\boldsymbol{\beta}|X, \mathbf{y}) = \prod_{i=1}^n \binom{n_i}{y_i} \left\{ F(\mathbf{x}_i' \boldsymbol{\beta}) \right\}^{y_i} \left\{ 1 - F(\mathbf{x}_i' \boldsymbol{\beta}) \right\}^{n_i - y_i}, \quad (2)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ and $X = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)'$ is the $n \times (k+1)$ design matrix. Throughout the article, we assume that $F(\cdot)$ is twice differentiable and that $f(z) = dF(z)/dz$ denotes the probability density function (pdf). Then Jeffreys's prior for $\boldsymbol{\beta}$ under the binomial regression model (1) is given by

$$\pi(\boldsymbol{\beta}|X) \propto |X'W(\boldsymbol{\beta})X|^{1/2}, \quad (3)$$

where $|X'W(\boldsymbol{\beta})X|$ denotes the determinant of the matrix $X'W(\boldsymbol{\beta})X$, $W(\boldsymbol{\beta}) = \text{diag}(w_1(\boldsymbol{\beta}), w_2(\boldsymbol{\beta}), \dots, w_n(\boldsymbol{\beta}))$, and

$$w_i(\boldsymbol{\beta}) = \frac{n_i \left\{ f(\mathbf{x}_i' \boldsymbol{\beta}) \right\}^2}{F(\mathbf{x}_i' \boldsymbol{\beta}) \left\{ 1 - F(\mathbf{x}_i' \boldsymbol{\beta}) \right\}} \quad (4)$$

for $i = 1, 2, \dots, n$. The Jeffreys prior given by (3) does not have a closed form in general, in the sense that it is known only up to a prior normalizing constant and is generally improper for most generalized linear models except for binary regression models, such as logistic, probit, and complementary log-log regression models, as shown by Ibrahim and Laud (1991). In addition, it has several additional attractive properties, which we formally state as follows.

Proposition 1

For the binomial regression model (1), assume that X is of full rank. Then Jeffreys's prior (3) for β is proper, and the corresponding moment-generating function of β exists.

Proposition 1 is a special case of theorem 2.1 of Ibrahim and Laud (1991).

Proposition 2

Assume that $F(z)$ is symmetric in the sense that $F(-z) = 1 - F(z)$ and $f(-z) = f(z)$. Then Jeffreys's prior $\pi(\beta|X)$ in (3) is symmetric about $\mathbf{0}$, that is, $\pi(-\beta|X) = \pi(\beta|X)$, $\forall \beta \in R^{k+1}$, where R^{k+1} denotes $(k+1)$ -dimensional Euclidean space.

The proof of Proposition 2 follows directly from the fact that $w_i(-\beta) = w_i(\beta)$ for $i = 1, 2, \dots, n$.

Theorem 1

Let $q(z) = \log \left[\frac{\{f(z)\}^2}{F(z)\{1-F(z)\}} \right] = 2 \log f(z) - \log F(z) - \log \{1-F(z)\}$. Assume that (a) X is of full rank; (b) $q(z)$ has a unique mode, z_{mod} ; and (c) $q'(z) < 0$ if $z > z_{mod}$, $q'(z_{mod}) = 0$, and $q'(z) > 0$ if $z < z_{mod}$. Then Jeffreys's prior $\pi(\beta|X)$ in (3) is unimodal, and its unique mode is $\beta_{mod} = (z_{mod}, 0, \dots, 0)'$.

Assumptions (b) and (c) in Theorem 1 are satisfied for several binomial regression models, including logistic, probit, and complementary log-log regressions. We formally state this result in the next theorem.

Theorem 2

Assumptions (b) and (c) in Theorem 1 hold for $F(z) = \exp(z)/\{1 + \exp(z)\}$, $F(z) = \Phi(z)$ [the $N(0, 1)$ cdf], and $F(z) = 1 - \exp\{-\exp(z)\}$, corresponding to logistic, probit, and complementary log-log regression models. Furthermore, Jeffreys's prior $\pi(\beta|X)$ has a unique mode $\beta_{mod} = \mathbf{0}$ for the logistic and probit regression models, and $\beta_{mod} = (.466, 0, \dots, 0)'$ for the complementary log-log regression model.

Our next result establishes an interesting property of Jeffreys's prior under logistic, probit, and complementary log-log regressions—that the tails of Jeffreys's prior, regardless of sample size, are always lighter than that of a multivariate t distribution. Toward this end, let $g(\beta|\Sigma, \nu)$ denote the pdf of a $(k+1)$ -dimensional multivariate t -distribution with ν degrees of freedom, location $\mathbf{0}$, and a positive definite dispersion matrix Σ , that is,

$$g(\beta|\Sigma, \nu) = \frac{\Gamma\{(\nu+k+1)/2\}}{\Gamma(\nu/2)(\nu\pi)^{(k+1)/2}} |\Sigma|^{-1/2} \times \left(1 + \frac{1}{\nu} \beta' \Sigma^{-1} \beta\right)^{-(\nu+k+1)/2}. \quad (5)$$

Then we are led to the following theorem, which characterizes the tail behavior of Jeffreys's prior.

Theorem 3

Assume that X is of full rank. Then Jeffreys's prior $\pi(\beta|X)$ in (3) under logistic regression, probit regression, and complementary log-log regression has lighter tails than $g(\beta|\Sigma, \nu)$ for any $\nu >$

0, that is, $\lim_{\|\beta\| \rightarrow \infty} \frac{\pi(\beta|X)}{g(\beta|\Sigma, \nu)} = 0$.

In the following theorem, we examine the tail behavior of Jeffreys's prior compared with a $(k + 1)$ -dimensional multivariate normal distribution.

Theorem 4

Let $\phi_{k+1}(\beta|\Sigma_N)$ denote the probability density function of the $(k + 1)$ -dimensional normal distribution $N_{k+1}(0, \Sigma_N)$, where Σ_N is a $(k + 1) \times (k + 1)$ positive definite matrix. Then the following results hold:

- a. Under logistic regression, we have $\lim_{\|\beta\| \rightarrow \infty} \pi(\beta|X) / \phi_{k+1}(\beta|\Sigma_N) = \infty$, which implies that Jeffreys's prior $\pi(\beta|X)$ under logistic regression always has heavier tails than the normal distribution, regardless of n .
- b. Let $X_{i_1 i_2 \dots i_{k+1}}^* = (\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \dots, \mathbf{x}_{i_{k+1}})'$ be a $(k + 1) \times (k + 1)$ submatrix of X . If there exists $(i_1, i_2, \dots, i_{k+1})$ such that $X_{i_1 i_2 \dots i_{k+1}}^*$ is of full rank and $\Sigma_N^{-1} - \frac{1}{2} (X_{i_1 i_2 \dots i_{k+1}}^*)' \times X_{i_1 i_2 \dots i_{k+1}}^* > 0$ (i.e., positive definite), then the normal distribution $N_{k+1}(0, \Sigma_N)$ has lighter tails than Jeffreys's prior $\pi(\beta|X)$ under probit regression. If $\Sigma_N^{-1} - \frac{1}{2} (X_{i_1 i_2 \dots i_{k+1}}^*)' \times X_{i_1 i_2 \dots i_{k+1}}^* < 0$ (i.e., negative definite) for all $(k + 1) \times (k + 1)$ full-rank submatrices $X_{i_1 i_2 \dots i_{k+1}}^*$ of X , then Jeffreys's prior $\pi(\beta|X)$ under probit regression has lighter tails than the normal distribution $N_{k+1}(0, \Sigma_N)$.
- c. Let $\beta = r\mathbf{d}$, where $r \geq 0$ and $\mathbf{d} = (d_0, d_1, d_2, \dots, d_k)'$ denotes a $(k + 1)$ -dimensional vector of unit direction such that $\|\mathbf{d}\| = \sqrt{\mathbf{d}'\mathbf{d}} = 1$. Under complementary log–log regression, Jeffreys's prior $\pi(\beta|X)$ has lighter tails than a $N_{k+1}(0, \Sigma_N)$ distribution in certain directions $\mathbf{d}$, including $\mathbf{d} = (1, 0, 0, \dots, 0)'$, and has heavier tails than a $N_{k+1}(0, \Sigma_N)$ distribution in some other directions $\mathbf{d}$, including $\mathbf{d} = (-1, 0, 0, \dots, 0)'$.

Next we characterize the conditional Jeffreys's prior distribution of β_0 .

Proposition 3

For Jeffreys's prior $\pi(\beta|X)$ given in (3) for general binomial regression, the conditional prior distribution of β_0 (the intercept), given $\beta_1 = \dots = \beta_k = 0$, is given by

$$\pi(\beta_0 | \beta_1 = \dots = \beta_k = 0, X) \propto \left[\frac{f^2(\beta_0)}{F(\beta_0) \{1 - F(\beta_0)\}} \right]^{(k+1)/2},$$

and the conditional posterior distribution of β_0 , given $\beta_1 = \dots = \beta_k = 0$, is given by

$$\begin{aligned} \pi(\beta_0 | \beta_1 = \dots = \beta_k = 0, X, \mathbf{y}) \\ \propto \{f(\beta_0)\}^{k+1} \{F(\beta_0)\}^{\sum_{i=1}^n y_i - (k+1)/2} \\ \times \{1 - F(\beta_0)\}^{\sum_{i=1}^n (n_i - y_i) - (k+1)/2}. \end{aligned}$$

The proof of Proposition 3 is straightforward. Note that the results given in Proposition 3 imply that the conditional Jeffreys's prior distribution of β_0 does not depend on the sample size n , but the conditional posterior does. This result sheds light on the asymptotic behavior of Jeffreys's prior—namely, that it does not converge to any well-known distribution as $n \rightarrow \infty$.

Let $C_0(X) = \int_{R^{k+1}} |X'W(\beta)X|^{1/2} d\beta$ and $C(X) = \int_{R^{k+1}} L(\beta|X, \mathbf{y}) |X'W(\beta)X|^{1/2} d\beta$, which correspond to the prior and posterior normalizing constants. We present an interesting result

—that the prior and posterior normalizing constants based on Jeffreys's prior are invariant under a one-to-one linear transformation of the covariates. Let V denote a $(k+1) \times (k+1)$ matrix. We are led to the following theorem.

Theorem 5

Assume that V is of full rank $(k+1)$. We have

$$C_0(X) = C_0(XV) \quad \text{and} \quad C(X) = C(XV). \quad (6)$$

Note that when V is of full rank, XV denotes a one-to-one linear transformation of the covariates. By taking $V = \text{diag}(1, v_1, v_2, \dots, v_k)$ with $v_1 > 0, v_2 > 0, \dots, v_k > 0$, (6) implies that $C_0(X)$ and $C(X)$ are scale-invariant with respect to the covariates. The scale-invariance of $C_0(X)$ and $C(X)$ in the covariates is a desirable property in Bayesian variable selection. Because the posterior model probability under a uniform prior on the model space is a function of prior and posterior normalizing constants, the result given in Theorem 5 also implies that the posterior model probability is scale-invariant in the covariates. In this sense Jeffreys's prior is as attractive as Zellner's g-prior (Zellner 1986), which also leads to posterior model probabilities that are scale invariant in the covariates.

Finally, we examine a theoretical connection between the Bayes information criterion (BIC) (Schwarz 1978) and the ratio of the posterior normalizing constant and the prior normalizing constant under Jeffreys's prior. Toward this end, we assume that there are only $k+1$ distinct rows in X , denoted by $\tilde{\mathbf{x}}'_j, j = 1, 2, \dots, k+1$. Under this assumption, we combine the binomial counts into $k+1$ aggregated counts corresponding to those $\tilde{\mathbf{x}}'_j$'s, and the aggregated likelihood function of (2) is given by

$$L(\beta|X_A, \mathbf{y}) = \left\{ \prod_{i=1}^n \binom{n_i}{y_i} \right\} \left[\prod_{j=1}^{k+1} \left\{ F(\tilde{\mathbf{x}}'_j \beta) \right\}^{y_{Aj}} \left\{ 1 - F(\tilde{\mathbf{x}}'_j \beta) \right\}^{n_{Aj} - y_{Aj}} \right], \quad (7)$$

where $X_A = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_{k+1})'$, $y_{Aj} = \sum_{i: \mathbf{x}_i = \tilde{\mathbf{x}}_j} y_i$, and $n_{Aj} = \sum_{i: \mathbf{x}_i = \tilde{\mathbf{x}}_j} n_i$. Using (7), Jeffreys's prior in (3) reduces to $\pi(\beta|X_A) \propto |X'_A W_A(\beta) X_A|^{1/2}$, where $W_A(\beta) = \text{diag}(w_{A1}(\beta), \dots, w_{A,k+1}(\beta))$ and

$$w_{Aj}(\beta) = \frac{n_{Aj} \left\{ f(\tilde{\mathbf{x}}'_j \beta) \right\}^2}{F(\tilde{\mathbf{x}}'_j \beta) \left\{ 1 - F(\tilde{\mathbf{x}}'_j \beta) \right\}}. \quad \text{It is easy to see that when } X_A \text{ is of full rank and } n_{Aj} \geq 1, \pi(\beta|$$

X_A) is proper. Let $p_j = p_j(\beta) = F(\tilde{\mathbf{x}}'_j \beta)$ for $j = 1, 2, \dots, k+1$. When X_A is of full rank, $p_1, p_2, \dots, p_{k+1}$ are $(k+1)$ unconstrained binomial probabilities. The aggregated likelihood given in (7) thus can be written as a function of $(p_1, p_2, \dots, p_{k+1})$ given by

$$L(p_1, p_2, \dots, p_{k+1}|X_A, \mathbf{y}) = \left\{ \prod_{i=1}^n \binom{n_i}{y_i} \right\} \left[\prod_{j=1}^{k+1} p_j^{y_{Aj}} (1 - p_j)^{n_{Aj} - y_{Aj}} \right]. \quad \text{In terms of the binomial probabilities, Jeffreys's prior and the corresponding posterior distribution have closed-form expressions. Specifically, Jeffreys's prior and the corresponding posterior distribution of } p_1, p_2, \dots, p_{k+1} \text{ are given by}$$

$$\begin{aligned}\pi(p_1, p_2, \dots, p_{k+1} | X_A) \\ = \left\{ B\left(\frac{1}{2}, \frac{1}{2}\right) \right\}^{-(k+1)} \prod_{j=1}^{k+1} \left\{ p_j (1 - p_j) \right\}^{-1/2}\end{aligned}\quad (8)$$

and

$$\begin{aligned}\pi(p_1, p_2, \dots, p_{k+1} | X_A, \mathbf{y}) \\ = \prod_{j=1}^{k+1} \frac{p_j^{y_{Aj}-1/2} (1-p_j)^{n_{Aj}-y_{Aj}-1/2}}{B\left(\frac{1}{2}+y_{Aj}, \frac{1}{2}+n_{Aj}-y_{Aj}\right)}.\end{aligned}\quad (9)$$

It is easy to show that

$$\begin{aligned}\pi(\beta | X_A) \\ = \pi(p_1, p_2, \dots, p_{k+1} | X_A) \left| \frac{(\partial p_1, p_2, \dots, \partial p_{k+1})}{\partial \beta} \right|,\end{aligned}\quad (10)$$

where $\frac{(\partial p_1, \partial p_2, \dots, \partial p_{k+1})}{\partial \beta} = \text{diag} \left(f'(\mathbf{x}_1 \beta), \dots, f'(\mathbf{x}_{k+1} \beta) \right) X_A$. Using (8) and (10), we obtain a closed-form expression of the normalizing constant for Jeffreys's prior, given by

$$\begin{aligned}C_0(X) &= \int_{R^{k+1}} |X'_A W_A(\beta) X_A|^{1/2} d\beta \\ &= \left(\prod_{j=1}^{k+1} n_{Aj} \right)^{1/2} \left\{ B\left(\frac{1}{2}, \frac{1}{2}\right) \right\}^{k+1} \\ &= \pi^{k+1} \left(\prod_{j=1}^{k+1} n_{Aj} \right)^{1/2},\end{aligned}\quad (11)$$

where $B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)}$ denotes the beta function. Similarly, using (9) and (10), we obtain the posterior normalizing constant based on Jeffreys's prior as

$$\begin{aligned}C(X) &= \int_{R^{k+1}} L(\beta | X_A, \mathbf{y}) |X'_A W_A(\beta) X_A|^{1/2} d\beta \\ &= \left\{ \prod_{i=1}^n \binom{n_i}{y_i} \right\} \left(\prod_{j=1}^{k+1} n_{Aj} \right)^{1/2} \\ &\quad \times \left\{ \prod_{j=1}^{k+1} B\left(\frac{1}{2}+y_{Aj}, \frac{1}{2}+n_{Aj}-y_{Aj}\right) \right\}.\end{aligned}\quad (12)$$

Let $N = \sum_{i=1}^n n_i$, $\hat{\alpha}_j = \frac{n_{Aj}}{N}$ and $\hat{\mu}_j = \frac{y_{Aj}}{N}$ for $j = 1, 2, \dots, k+1$. Also, let $\hat{\beta}$ denote the maximum likelihood estimate of β . Then, using (7), the BIC is given by

$$\begin{aligned}
\text{BIC} &= -2 \log L(\widehat{\beta}|X_{A,y}) + (k+1) \log(N) \\
&= -2 \log \left\{ \prod_{i=1}^n \binom{n_i}{y_i} \right\} - 2 \sum_{j=1}^{k+1} \left\{ y_{Aj} \log \left(\frac{y_{Aj}}{n_{Aj}} \right) \right. \\
&\quad \left. + (n_{Aj} - y_{Aj}) \log \left(\frac{n_{Aj} - y_{Aj}}{n_{Aj}} \right) \right\} \\
&\quad + (k+1) \log(N).
\end{aligned} \tag{13}$$

The following theorem characterizes the connection between the BIC and the normalizing constants.

Theorem 6

Assume that (a) $\lim_{N \rightarrow \infty} \widehat{\alpha}_j = \alpha_j$ and $\lim_{N \rightarrow \infty} \widehat{\mu}_j = \mu_j$ exist and (b) $0 < \alpha_j < 1$ and $0 < \mu_j < \alpha_j$ for $j = 1, 2, \dots, k+1$. Then for large N , we have

$$\begin{aligned}
-2 \{ \log C(X) - \log C_0(X) \} \\
= \text{BIC} + \sum_{j=1}^{k+1} \log \left(\frac{\pi \widehat{\alpha}_j}{2} \right) + o\left(\frac{1}{N}\right),
\end{aligned} \tag{14}$$

where $C_0(X)$, $C(X)$, and BIC are defined by (11), (12), and (13).

The proof of Theorem 6 follows directly from Stirling's formula, and we omit the details for brevity. The theoretical connection given in (14) is quite interesting. First, $-2 \{ \log C(X) - \log C_0(X) \}$ acts quite similarly to the BIC. Second, in addition to the dimensional penalty of $(k+1) \log N$ in the BIC, the dimensional penalty term in $-2 \{ \log C(X) - \log C_0(X) \}$ also depends on $(\alpha_1, \alpha_2, \dots, \alpha_{k+1})$, which describes the “joint distribution” of the covariate $\mathbf{x}_i$, which takes on the $k+1$ distinct values $\tilde{\mathbf{x}}_j, j = 1, \dots, k+1$. Thus Bayesian variable selection based on Jeffreys's prior can yield very different results than the BIC, especially when N is small.

3. COMPUTATIONAL IMPLEMENTATION OF JEFFREYS'S PRIOR

As shown in Theorem 3, under logistic regression, probit regression, and complementary log-log regression, Jeffreys's prior $\pi(\beta|X)$ in (3) has lighter tails than $g(\beta|\Sigma, \nu)$ in (5) for any $\nu > 0$. In addition, the likelihood $L(\beta|X, \mathbf{y})$ is bounded above, so it follows that the posterior has lighter tails than the t -distribution. Because of Theorems 3 and 4, we can completely avoid Gibbs sampling or other computationally intensive MCMC schemes for carrying out posterior computations. Toward this goal, we propose an importance sampling approach (see, e.g., Geweke 1989) to compute the prior and posterior quantities of interest. To obtain appropriate importance sampling densities for Jeffreys's prior and the corresponding posterior distribution, we consider a more general form of the multivariate t -distribution with density

$$\begin{aligned}
g(\beta|\mu, \Sigma, \nu) \\
= \frac{\Gamma((\nu+k+1)/2)}{\Gamma(\nu/2)(\nu\pi)^{(k+1)/2}} |\Sigma|^{-1/2} \\
\times \left(1 + \frac{1}{\nu} (\beta - \mu)' \Sigma^{-1} (\beta - \mu) \right)^{-(\nu+k+1)/2}.
\end{aligned} \tag{15}$$

For computing the prior normalizing constant, we specify $\mu = \beta_{mod}$ in (15), because Jeffreys's prior has a unique mode at β_{mod} as shown in Theorem 1. To determine Σ in (15), we match the Hessians (curvatures) of Jeffreys's prior and the t -distribution at β_{mod} as

$$\kappa_0 \left. \frac{\partial^2 \log \pi(\beta|X)}{\partial \beta \partial \beta'} \right|_{\beta=\beta_{mod}} = \left. \frac{\partial^2 \log g(\beta|\mu=\beta_{mod}, \Sigma, \nu)}{\partial \beta \partial \beta'} \right|_{\beta=\beta_{mod}}, \quad (16)$$

where $\pi(\beta|X)$ and $g(\beta|\mu, \Sigma, \nu)$ are given by (3) and (15), and $\kappa_0 > 0$ is a fixed dispersion adjustment parameter. We follow a similar strategy for computing the posterior normalizing constant. We match the mode and the Hessian (curvature) at the mode of the posterior distribution μ and Σ in the importance sampling density (15). In doing so, we are led to $\mu=\hat{\mu}=\arg \max_{\beta \in R^{k+1}} \{\log [L(\beta|X, \mathbf{y}) \pi(\beta|X)]\}$ and

$$\Sigma^{-1} = -\kappa_1 \frac{\nu}{\nu+k+1} \left. \frac{\partial^2 \log L(\beta|X, \mathbf{y}) \pi(\beta|X)}{\partial \beta \partial \beta'} \right|_{\beta=\hat{\mu}}, \quad (17)$$

where $\kappa_1 > 0$ is a fixed dispersion adjustment parameter. A detailed discussion of the specification of ν and κ_0 (or κ_1) in determining the importance sampling density $g(\beta|\mu, \Sigma, \nu)$ is given in Appendix B of the supplementary document.

The steps of the importance sampling method are quite simple. Let $\mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ denote the gamma distribution with shape parameter $\nu/2$ and scale parameter $\nu/2$. Then the importance sampling algorithm for computing the prior normalizing constant $C_0 = C_0(X)$ can be stated as follows:

Step 1: Generate a random sample $\{\beta_1, \beta_2, \dots, \beta_Q$ of size Q from $g(\beta|\mu = \beta_{mod}, \Sigma, \nu)$,

where for each q , independently generate $\lambda_q \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ and $\beta_q \sim N_{k+1}(\beta_{mod}, \Sigma/\lambda_q)$.

Step 2: Compute the Monte Carlo estimate of the prior normalizing constant C_0 as

$$\widehat{C}_0 = \frac{1}{Q} \sum_{q=1}^Q \frac{|X'W(\beta_q)X|^{1/2}}{g(\beta_q|\mu=\beta_{mod}, \Sigma, \nu)}.$$

In step 2 we also may calculate $\log(\widehat{C}_0)$ instead of $\widehat{C}_0$ for greater stability and accuracy. In addition, we compute the relative Monte Carlo (MC) standard error (SE) as

$$\text{RSE}(\widehat{C}_0) = \frac{1}{\widehat{C}_0} \left\{ \frac{1}{Q(Q-1)} \sum_{q=1}^Q \left[\frac{|X'W(\beta_q)X|^{1/2}}{g(\beta_q|\mu=\beta_{mod}, \Sigma, \nu)} - \widehat{C}_0 \right]^2 \right\}^{1/2}. \quad \text{Note that by the standard delta method, we can show that } \text{RSE}(\widehat{C}_0) \text{ is indeed the MC SE of } \log(\widehat{C}_0). \text{ In practice, we recommend choosing } Q \text{ and } \kappa_0 \text{ so that } \text{RSE}(\widehat{C}_0) \text{ is approximately .01 or less. The importance sampling algorithm for computing the posterior normalizing constant } C = C(X) \text{ is analogous to that for the prior normalizing constant.}$$

To examine the performance of the proposed importance sampling algorithm, we consider the logistic regression model with a single binary covariate. We generate a simulated data set of

size $n = 100$ with $x_{i1} \sim \text{Bernoulli}(.6)$ and $y_i|x_{i1} \sim \text{Bernoulli}(p_i)$, where $p_i = \frac{\exp(\mathbf{x}_i'\beta)}{1+\exp(\mathbf{x}_i'\beta)}$, $\mathbf{x}_i = (1, x_{i1})'$, and $\beta = (.5, 1.0)'$. We implemented the proposed importance sampling algorithm with various values of κ_0 and κ_1 . The results are given in Table 1. The prior and posterior normalizing

constants in this example involve a two-dimensional integral. As discussed in Appendix B, $\beta_{mod} = \mathbf{0}$ and $v = 3.37$ is a guide value for v in $g(\beta|\mu = \mathbf{0}, \Sigma, v)$ for computing the prior normalizing constant under logistic regression. From Table 1, we see that both $v = 3.37$ and $v = 5$ work well for computing both the prior and posterior normalizing constants. The MC SEs are all smaller than .01 with $Q = 10,000$. Moreover, we see that $\log(\widehat{C}_0)$, $\log(\widehat{C}_1)$, and the MC SE are extremely consistent and robust for several values of v , κ_0 , and κ_1 . Moreover, the MC SE for $v = 3.37$ is smallest among all values of v , confirming our theoretical results concerning the choice of v .

Next we discuss how to compute the posterior estimates of β under Jeffreys's prior via the proposed importance sampling method. Let $\{\beta_1, \beta_2, \dots, \beta_Q\}$, where $\beta_q = (\beta_{q0}, \beta_{q1}, \dots, \beta_{qk})'$, $q = 1, 2, \dots, Q$, be a random sample of size Q from $g(\beta|\mu, \Sigma, v)$. Then an MC estimate of

the h th posterior moment of β_j is given by
$$\widehat{E}(\beta_j^h|X, y) = \frac{\sum_{q=1}^Q \beta_{qj}^h \omega_q}{\sum_{q=1}^Q \omega_q}, \text{ where } \omega_q = \frac{\pi^*(\beta_q|X, y)}{g(\beta_q|\mu, \Sigma, v)}$$
 for $q = 1, 2, \dots, Q$. Using $\widehat{E}(\beta_j^h|X, y)$, we can easily compute the posterior mean and standard deviation of β_j . To compute the highest posterior density (HPD) interval of β_j through the importance sampling method, we use the MC method proposed by Chen and Shao (1999).

Specifically, for $0 \leq \gamma < 1$, define $\widehat{\beta}_j^{(\gamma)} = \beta_{j(1)}$ if $\gamma = 0$ and $\beta_{j(q)}$ if $\sum_{l=1}^{q-1} \omega_l < \gamma \leq \sum_{l=1}^q \omega_l$, where $\beta_{j(q)}$ is the q th smallest of $\{\beta_{j(l)}, l = 1, 2, \dots, Q\}$. To obtain a $100(1 - \alpha)\%$ HPD interval for β_j , we let $R_q(Q) = (\widehat{\beta}_j^{(q/Q)}, \widehat{\beta}_j^{((q+(1-\alpha)Q)/Q)})$ for $q = 1, 2, \dots, Q - [(1 - \alpha)Q]$, where $[(1 - \alpha)Q]$ denotes the integer part of $(1 - \alpha)Q$. Then the $100(1 - \alpha)\%$ HPD interval is $R_{q^*}(Q)$, the interval with the smallest width among all $R_q(Q)$'s.

4. ANALYSIS OF PROSTATE CANCER DATA

To further motivate the proposed methodology, we consider data from a retrospective cohort study of men treated with radical prostatectomy ($n = 968$) between 1988 and 2000, which is a subset of the data published by D'Amico et al. (2002). The primary endpoint that D'Amico et al. (2002) considered was prostate-specific antigen (PSA) recurrence free survival. In our analysis, we consider Pathological Extracapsular Extension (PECE) a binary response variable (y) that takes values 0 and 1, where 1 denotes that the cancer has penetrated the prostate wall and 0 indicates otherwise. We consider five prognostic factors: age, natural logarithm of PSA (LogPSA), percent positive prostate biopsies (ppb), biopsy Gleason score, and the 1992 American Joint Commission on Cancer (AJCC) clinical tumor category. The covariates age, LogPSA, and ppb are continuous. We dichotomize biopsy Gleason score as GS7 and GS8H, where (GS7, GS8H) takes values (0, 0), (1, 0), and (0, 1), corresponding to biopsy Gleason scores of ≤ 6 , 7, and 8–10. Similarly, we dichotomize the clinical tumor category as T2b and T2c, where (T2b, T2c) takes values (0, 0), (1, 0), and (0, 1), corresponding to the clinical tumor categories T1c or T2a, T2b, and T2c.

First, we carried out variable selection. For all of 32 possible models, including an intercept with five covariates, we computed the Akaike information criterion (AIC), the BIC, and the posterior model probabilities based on Jeffreys's prior. To compute the posterior model probabilities based on Jeffreys's prior, we implemented the importance sampling algorithm proposed in Section 3 with an MC sample size of $Q = 20,000$, $\kappa_0 = 1$, $\kappa_1 = 2$, and $v = 3.37$ and $v = 5$ for the prior and posterior normalizing constants. The best model under the BIC and the model with the highest posterior probability based on Jeffreys's prior is (LogPSA, ppb, GS7, GS8H). For this best model, the BIC value is 925.627, and the posterior probability is .871. The second-best model under these two criteria is (age, LogPSA, ppb, GS7, GS8H), with a BIC value of 928.574 and posterior probability of .119. The AIC selects the full model (age, LogPSA, ppb, GS7, GS8H, T2b, T2c) as the best model. The AIC values are 899.168 for the

full model and 901.251 for the best BIC model. We note that the AIC gives different results than the BIC and the highest posterior probability model, apparently with no good scientific reason.

Next, we computed the posterior estimates under Jeffreys's prior for the highest posterior probability model, (LogPSA, ppb, GS7, GS8H), using the MC method with a MC sample of size $Q = 20,000$, as discussed in Section 3. Table 2 gives the posterior means (Estimates), the posterior standard deviations (SDs), and 95% HPD intervals for the β_j 's, along with the corresponding maximum likelihood (ML) estimates, standard errors (SEs), and p values. We can see from Table 2 that the posterior estimates based on Jeffreys's prior are very close to the ML estimates, which is intuitively appealing, thus demonstrating that Jeffreys's prior is indeed noninformative. We also can see from this table that all selected variables are highly significant, which implies that these are the most important prognostic factors in predicting the binary response PECE. We note that under the full model, age, T2b, and T2c all have 95% HPD intervals that contain 0.

5. DISCUSSION

We have examined theoretical and computational properties of Jeffreys's prior for inference in binomial regression models. We have derived several theoretical properties of Jeffreys's prior for binomial regression models, including unimodality, symmetry, mode, and tail behavior. For variable selection problems, we have established a theoretical connection between the BIC and the ratio of the posterior and prior normalizing constants with Jeffreys's prior under binomial regression with a general link. As shown in Theorem 6, the dimension penalty under Jeffreys's prior depends on the covariates as well as on the penalty term imposed by the BIC. The penalty term for the BIC depends only on k and N , and thus the BIC does not have a penalty as sophisticated as that of Jeffreys's prior for variable selection problems. This is an interesting theoretical difference that may have an impact in variable selection problems. Future work includes investigating the performance of Jeffreys's prior in variable selection and carrying out computations for variable selection with Jeffreys's prior for high-dimensional problems.

Acknowledgements

The authors thank the editor, the associate editor, and three referees for helpful comments and suggestions that have improved the article. Dr. Chen's and Dr. Ibrahim's research was supported in part by National Institute of Health grants GM 70335 and CA 74015.

REFERENCES

- Berger, JO. Statistical Decision Theory and Bayesian Analysis. Vol. 2nd ed.. Springer-Verlag; New York: 1985.
- Berger JO. Bayesian Analysis: A Look at Today and Thoughts of Tomorrow. *Journal of the American Statistical Association* 2000;95:1269–1127.
- Berger JO. The Case for Objective Bayesian Analysis. *Bayesian Analysis* 2006;1:385–402.
- Box, GEP.; Tiao, GC. *Bayesian Inference in Statistical Analysis*. Addison-Wesley; Reading, MA: 1973.
- Chen M-H, Shao Q-M. Monte Carlo Estimation of Bayesian Credible and HPD Intervals. *Journal of Computational and Graphical Statistics* 1999;8:69–92.
- D'Amico AV, Whittington R, Malkowicz SB, Cote K, Loffredo M, Schultz D, Chen M-H, Tomaszewski JE, Renshaw AA, Wein A, Richie JP. Biochemical Outcome Following Radical Prostatectomy or External Beam Radiation Therapy for Clinically Localized Prostate Cancer in the PSA Era. *Cancer* 2002;95:281–286. [PubMed: 12124827]
- Geweke J. Bayesian Inference in Econometrics Models Using Monte Carlo Integration. *Econometrica* 1989;57:1317–1340.

- Ibrahim JG, Laud PW. On Bayesian Analysis of Generalized Linear Models Using Jeffreys's Prior. *Journal of the American Statistical Association* 1991;86:981–986.
- Jeffreys H. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London* 1946;196:453–461.Ser. A
- Jeffreys, H. *Theory of Probability*. Vol. 3rd ed.. Oxford University Press; Oxford, U.K.: 1961.
- Kass RE. The Geometry of Asymptotic Inference. *Statistical Science* 1989;4:188–234.(with discussion)
- Kass RE. Data Translated Likelihood and Jeffreys's Rules. *Biometrika* 1990;77:107–114.
- Kass RE, Wasserman L. The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* 1996;91:1343–1370.
- Schwarz G. Estimating the Dimension of a Model. *The Annals of Statistics* 1978;6:461–464.
- Zellner, A. On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions. In: Goel, P.; Zellner, A., editors. *Bayesian Inference and Decision Techniques*. Elsevier Science; Amsterdam: 1986. p. 233-243.

Table 1
Monte Carlo estimates of log C₀ and log C with MC size Q̄ = 10,000

v	Jeffrey's prior			Posterior		
	k ₀	log Ĉ ₀	MC SE	k ₁	log Ĉ	MC SE
1	1	6.137	.008	2	-56.907	.007
3.37		6.131	.002		-56.900	.003
5		6.133	.002		-56.895	.003
20		6.139	.012		-56.881	.007
3.37	.5	6.144	.005	1	-56.884	.006
	2	6.127	.005	3	-56.906	.004
5	.5	6.129	.004	1	-56.898	.004
	2	6.139	.008	3	-56.890	.005
True value		log C ₀ = 6.132			log C = -56.890	

Table 2

Posterior estimates of β for the prostate cancer data

Variable	ML estimates		Posterior estimates		
	Estimate	SE	p value	Estimate	SD
Intercept	−3.895	.304	<.0001	−3.896	.307
LogPSA	.696	.135	<.0001	.696	.135
ppb	2.376	.355	<.0001	2.376	.356
GS7	.705	.182	.0001	.706	.182
GS8H	1.420	.337	<.0001	1.420	.337
				95% HPD interval	
				(−4.586, −3.222)	
				(.400, 1.004)	
				(1.612, 3.201)	
				(.283, 1.098)	
				(.639, 2.156)	