Objective Bayesian Variable Selection for Binomial Regression Models with Jeffreys's Prior

Ming-Hui Chen
Department of Statistics, University of Connecticut

This is a joint work with Joseph G. Ibrahim and Sungduk Kim
Presented at Bayesian Model Selection and Objective Methods,
University of Florida, Department of Statistics
January 11-12, 2008

Bayesian Variable Selection

- In the Bayesian context, fully Bayesian variable selection involves proper prior elicitation for all of the parameters arising from the various submodels in the model space.

Bayesian Variable Selection

- In the Bayesian context, fully Bayesian variable selection involves proper prior elicitation for all of the parameters arising from the various submodels in the model space.
- It requires numerical computation of the posterior model probabilities (or Bayes factors) for all of the submodels.

Bayesian Variable Selection

- In the Bayesian context, fully Bayesian variable selection involves proper prior elicitation for all of the parameters arising from the various submodels in the model space.
- It requires numerical computation of the posterior model probabilities (or Bayes factors) for all of the submodels.
- As is well known, for posterior model probabilities to be well defined, one needs to define proper priors for all of the model parameters arising from all of the submodels in the model space.

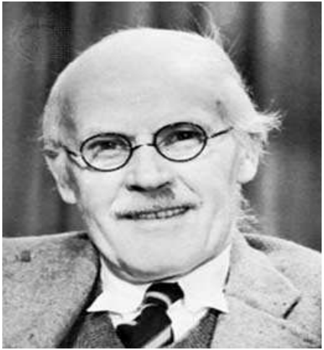
Bayesian Variable Selection

- In the Bayesian context, fully Bayesian variable selection involves proper prior elicitation for all of the parameters arising from the various submodels in the model space.
- It requires numerical computation of the posterior model probabilities (or Bayes factors) for all of the submodels.
- As is well known, for posterior model probabilities to be well defined, one needs to define proper priors for all of the model parameters arising from all of the submodels in the model space.
- This leads to the issue of specifying proper priors that are sufficiently noninformative so that the data can drive the inference, as is desired in most variable selection problems.

Bayesian Variable Selection

- In the Bayesian context, fully Bayesian variable selection involves proper prior elicitation for all of the parameters arising from the various submodels in the model space.
- It requires numerical computation of the posterior model probabilities (or Bayes factors) for all of the submodels.
- As is well known, for posterior model probabilities to be well defined, one needs to define proper priors for all of the model parameters arising from all of the submodels in the model space.
- This leads to the issue of specifying proper priors that are sufficiently noninformative so that the data can drive the inference, as is desired in most variable selection problems.
- Thus, in these types of problems, it becomes extremely attractive to have "semiautomatic" priors that are proper and require minimal elicitation.

Who is Harold Jeffreys?



Harold Jeffreys: April 22, 1891 - March 18, 1989

Who is Harold Jeffreys?

- Sir Harold Jeffreys is a British Astronomer and Geophysicist.

Who is Harold Jeffreys?

- Sir Harold Jeffreys is a British Astronomer and Geophysicist.
- As a statistician, he re-established the statistical theory of his time on Bayesian foundations.

Who is Harold Jeffreys?

- Sir Harold Jeffreys is a British Astronomer and Geophysicist.
- As a statistician, he re-established the statistical theory of his time on Bayesian foundations.
- His classical book is *Theory of Probability*, Third Edition, Oxford: Oxford University Press, 1961.

Jeffreys's Prior

- Jeffreys's prior is perhaps the most widely used noninformative prior in Bayesian analysis.

Jeffreys's Prior

- Jeffreys's prior is perhaps the most widely used noninformative prior in Bayesian analysis.
- In the context of binomial regression, Jeffreys's prior is proper for this model under very mild conditions (see Ibrahim and Laud, 1991)

Jeffreys's Prior

- Jeffreys's prior is perhaps the most widely used noninformative prior in Bayesian analysis.
- In the context of binomial regression, Jeffreys's prior is proper for this model under very mild conditions (see Ibrahim and Laud, 1991)
- Jeffrey's prior is simply the determinant of the square root of the Fisher information matrix.

Literature on Jeffrey's prior

- 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.

Literature on Jeffrey's prior

- 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.
- Two excellent books discussing Jeffreys's prior include Box and Tiao (1973) and Berger (1985).

Literature on Jeffrey's prior

Other relevant key references include Jeffreys (1946, 1961), Bernardo (1979), Eaves (1983), Kass (1989, 1990), Ibrahim and Laud (1991), Ye and Berger (1991), Berger and Bernardo (1989, 1992), McCulloch and Rossi (1992), Firth (1993), Mallick and Gelfand (1994), Gelfand and Mallick (1995), Kass and Raftery (1995), Raftery (1996), Kass and Wasserman (1996), Daniels (1999), Natarajan and Kass (2000), Berger, De Olivera, and Sansó (2001), Berger (2000, 2006), and Komaki (2006).

Unknown Properties of Jeffrey's prior

- What are the potential connections to normal or t distributions?

Unknown Properties of Jeffrey's prior

- What are the potential connections to normal or t distributions?
- What are the tail behavior of Jeffreys's prior, unimodality and symmetry properties?

Unknown Properties of Jeffrey's prior

- What are the potential connections to normal or t distributions?
- What are the tail behavior of Jeffreys's prior, unimodality and symmetry properties?
- What are techniques for sampling from Jeffreys's prior?

Unknown Properties of Jeffrey's prior

- What are the potential connections to normal or t distributions?
- What are the tail behavior of Jeffreys's prior, unimodality and symmetry properties?
- What are techniques for sampling from Jeffreys's prior?
- How does it perform in variable selection problems?

Logistic Regression Model

- Suppose that $\{(\mathbf{x}_i, y_i, n_i),\ i = 1, 2, \ldots, n\}$ are independent observations

Logistic Regression Model

- Suppose that $\{(\mathbf{x}_i, y_i, n_i),\ i = 1, 2, \ldots, n\}$ are independent observations
- y_i is the binomial response variable taking a value between 0 and n_i (≥ 1)

Logistic Regression Model

- Suppose that $\{(\mathbf{x}_i, y_i, n_i),\ i = 1, 2, \ldots, n\}$ are independent observations
- y_i is the binomial response variable taking a value between 0 and n_i (≥ 1)
- $\mathbf{x}_i = (1, x_{i1}, \cdots, x_{ik})'$ is a $(k+1) \times 1$ random vector of covariates.

Logistic Regression Model

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

$$f(y_i|x_i, n_i, \boldsymbol{\beta}) = \binom{n_i}{y_i}[F(\mathbf{x}_i'\boldsymbol{\beta})]^{y_i}[1 - F(\mathbf{x}_i'\boldsymbol{\beta})]^{n_i - y_i},\ i = 1, 2, \ldots, n,$$

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

Binomial Regression Model

- We assume throughout that $F(\cdot)$ is twice differentiable and $f(z) = dF(z)/dz$ denotes the probability density function (pdf).

Binomial Regression Model

- We assume throughout that $F(\cdot)$ is twice differentiable and $f(z) = dF(z)/dz$ denotes the probability density function (pdf).
- The likelihood function of β is

$$L(\beta|X, \mathbf{y}) = \prod_{i=1}^{n} \binom{n_i}{y_i} [F(\mathbf{x}_i'\beta)]^{y_i} [1 - F(\mathbf{x}_i'\beta)]^{n_i - y_i},$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ and $X = (\mathbf{x}_1, x_2, \ldots, x_n)'$ is the $n \times (k+1)$ design matrix.

The Jeffreys's Prior

The Jeffreys's prior for β under the logistic regression model is given by

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

where $|X'W(\beta)X|$ denotes the determinant of the matrix $X'WX$,

$$W(\beta) = \text{diag}(w_1(\beta), w_2(\beta), \ldots, w_n(\beta)),$$

and

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

for $i = 1, 2, \ldots, n$.

Useful Propositions

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

Useful Propositions

- **Proposition 1**: *For the binomial regression model (**??**), assume that X is of full rank. Then the Jeffreys's prior (1) for $\boldsymbol{\beta}$ is proper and the corresponding moment generating function of $\boldsymbol{\beta}$ exists.*
- **Proposition 2**: *Assume that $F(z)$ is symmetric in the sense that $F(-z) = 1 - F(z)$ and $f(-z) = f(z)$. Then, the Jeffreys's prior $\pi(\boldsymbol{\beta}|X)$ in (1) is symmetric about $\mathbf{0}$, i.e.,*

$$\pi(-\boldsymbol{\beta}|X) = \pi(\boldsymbol{\beta}|X) \ \forall \, \boldsymbol{\beta} \in R^{k+1},$$

where R^{k+1} denotes the $(k+1)$-dimensional Euclidean space.

Four Key Theorems

- Let

$$q(z) = \log\left[\frac{\{f(z)\}^2}{F(z)\{1-F(z)\}}\right] = 2\log f(z) - \log F(z) - \log\{1-F(z)\}.$$

Four Key Theorems

- Let

$$q(z) = \log\left[\frac{\{f(z)\}^2}{F(z)\{1 - F(z)\}}\right] = 2\log f(z) - \log F(z) - \log\{1 - F(z)\}.$$

- **Theorem 1**: *Assume that (i) X is full rank, (ii) $q(z)$ has a unique mode z_{mod}, and (iii) $q'(z) < 0$ if $z > z_{mod}$, $q'(z_{mod}) = 0$, and $q'(z) > 0$ if $z < z_{mod}$. Then the Jeffreys's prior $\pi(\beta|X)$ in (1) is unimodal and its unique mode is $\beta_{mod} = (z_{mod}, 0, \ldots, 0)'$.*

Four Key Theorems

Theorem 2: *The assumptions (ii) and (iii) 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 regressions, respectively. Furthermore, the Jeffreys's prior* $\pi(\boldsymbol{\beta}|X)$ *has unique mode* $\boldsymbol{\beta}_{mod} = \mathbf{0}$ *for logistic and probit regression models and* $\boldsymbol{\beta}_{mod} = (0.466, 0, \ldots, 0)'$ *for complementary log-log regression model.*

Four Key Theorems

- Let $g(\boldsymbol{\beta}|\Sigma,\nu)$ denote the pdf of a $(k+1)$-dimensional multivariate t-distribution defined by

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

Four Key Theorems

- Let $g(\boldsymbol{\beta}|\Sigma,\nu)$ denote the pdf of a $(k+1)$-dimensional multivariate t-distribution defined by

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

- **Theorem 3**: *Assume that X is of full rank. Assume that X is of full rank. Then, the Jeffreys's prior $\pi(\boldsymbol{\beta}|X)$ in (1) under logistic regerssion, probit regression, and complementary log-log regressions has lighter tails than $g(\boldsymbol{\beta}|\Sigma,\nu)$ for any $\nu > 0$, that is,*

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

Four Key Theorems

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.*
(i) Under logistic regression, we have

$$\lim_{||\beta||\to\infty} \frac{\pi(\beta|X)}{\phi_{k+1}(\beta|\Sigma_N)} = \infty,$$

which implies that the Jeffreys's prior $\pi(\beta|X)$ under logistic regression always has heavier tails than the normal distribution, regardless of n.

Four Key Theorems

Theorem 4 (continued):
(ii) Let $X^*_{i_1 i_2 \ldots i_{k+1}} = (\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_{k+1}})'$ *be a* $(k+1) \times (k+1)$ *submatrix of* X*. If there exists* $(i_1, i_2, \ldots, i_{k+1})$ *such that* $X^*_{i_1 i_2 \ldots i_{k+1}}$ *is full rank and* $\Sigma_N^{-1} - \frac{1}{2}(X^*_{i_1 i_2 \ldots i_{k+1}})' X^*_{i_1 i_2 \ldots i_{k+1}} > 0$ *(i.e., positively defenite), then the normal distribution* $N_{k+1}(0, \Sigma_N)$ *has lighter tails than the Jeffreys's prior* $\pi(\beta|X)$ *under probit regression. If* $\Sigma_N^{-1} - \frac{1}{2}(X^*_{i_1 i_2 \ldots i_{k+1}})' X^*_{i_1 i_2 \ldots i_{k+1}} < 0$ *(i.e., negatively definite) for all* $(k+1) \times (k+1)$ *full rank submatrices* $X^*_{i_1 i_2 \ldots i_{k+1}}$ *of* X*, the Jeffreys's prior* $\pi(\beta|X)$ *under probit regression has lighter tails than the normal distribution* $N_{k+1}(0, \Sigma_N)$*.*

Four Key Theorems

Theorem 4 (continued):
(iii) Let $\beta = r\mathbf{d}$, *where* $r \geq 0$ *and* $\mathbf{d} = (d_0, d_1, d_2, \ldots, d_k)'$ *denotes a* $(k+1)$*-dimensional vector of the unit direction such that* $||\mathbf{d}|| = \sqrt{\mathbf{d}'\mathbf{d}} = 1$*. Under complementary log-log regression, the Jeffreys's prior* $\pi(\beta|X)$ *has lighter tails than* $N_{k+1}(0, \Sigma_N)$ *in certain directions* $\mathbf{d}$ *such as* $\mathbf{d} = (1, 0, 0, \ldots, 0)'$ *and heavier tails than* $N_{k+1}(0, \Sigma_N)$ *in some other directions* $\mathbf{d}$ *such as* $\mathbf{d} = (-1, 0, 0, \ldots, 0)'$.

Four Key Theorems

Proposition 3:
For Jeffreys's prior $\pi(\boldsymbol{\beta}|X)$ given in (1) for general binomial regression, the conditional prior distribution of β_0 (the intercept) given $\beta_1 = \cdots = \beta_k = 0$ is given by

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

and the conditional posterior distribution of β_0 given $\beta_1 = \cdots = \beta_k = 0$ is given by $\pi(\beta_0|\beta_1 = \cdots = \beta_k = 0, X, \mathbf{y}) \propto \{f(\beta_0)\}^{k+1}\{F(\beta_0)\}^{\sum_{i=1}^n y_i - \frac{k+1}{2}}\{1 - F(\beta_0)\}^{\sum_{i=1}^n (n_i - y_i) - \frac{k+1}{2}}$.

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.

Bayesian Variable Selection with Jeffreys's Prior

- Since Jeffreys's prior is proper for the binomial model, it can therefore be considered as the default prior in computing posterior model probabilities.

Bayesian Variable Selection with Jeffreys's Prior

- Since Jeffreys's prior is proper for the binomial model, it can therefore be considered as the default prior in computing posterior model probabilities.
- As the dimension of a submodel in the model space varies from one model to another, Jeffreys's prior adjusts the dimensionality in an automatic fashion.

Bayesian Variable Selection with Jeffreys's Prior

- Since Jeffreys's prior is proper for the binomial model, it can therefore be considered as the default prior in computing posterior model probabilities.
- As the dimension of a submodel in the model space varies from one model to another, Jeffreys's prior adjusts the dimensionality in an automatic fashion.
- Since Jeffreys's prior is a noninformative prior, it leads to "objective" Bayesian variable selection as discussed in Bernardo (1979) and Berger and Bernardo (1992).

Bayesian Variable Selection with Jeffreys's Prior

Let $\mathcal{M}$ denote the model space. We enumerate the models in $\mathcal{M}$ by $m = 1, 2, \ldots, \mathcal{K}$, where $\mathcal{K} = 2^k$ is the dimension of $\mathcal{M}$ and model $\mathcal{K}$ denotes the full model.
Under model m, the likelihood function is given by

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

where $X^{(m)} = \left(\mathbf{x}_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)}\right)'$ is the $n \times k_m$ design matrix.
The corresponding Jeffreys's prior for $\beta^{(m)}$ is given by

$$\pi(\boldsymbol{\beta}^{(m)}|X^{(m)}, m) \propto \left|(X^{(m)})'W^{(m)}(\boldsymbol{\beta}^{(m)})X^{(m)}\right|^{1/2}.$$

Bayesian Variable Selection with Jeffreys's Prior

Let

$$C_{0m} = \int_{R^{k_m}} \left| (X^{(m)})' W^{(m)}(\beta^{(m)}) X^{(m)} \right|^{1/2} d\beta^{(m)}$$

and

$$C_m = \int_{R^{k_m}} L(\beta^{(m)} | X^{(m)}, \mathbf{y}, m) \left| (X^{(m)})' W^{(m)}(\beta^{(m)}) X^{(m)} \right|^{1/2} d\beta^{(m)}.$$

Suppose that we take a uniform prior on the model space $\mathcal{M}$, that is, the prior probability of model m is $p(m) = \frac{1}{\mathcal{K}}$ for $m \in \mathcal{M}$. Let $D = (\mathbf{y}, X)$ denote the observed data. Then, by Bayes theorem, the posterior probability of model m given the observed data D is given by

$$p(m|D) = \frac{C_m / C_{0m}}{\sum_{m^*=1}^{\mathcal{K}} C_{m^*} / C_{0m^*}}. \quad (2)$$

Model choice is then based on selecting the model which yields the largest posterior model probability $p(m|D)$.

Bayesian Variable Selection with Jeffreys's Prior

Let $\tilde{\mathbf{x}}_j^{(m)} = (x_{1j}^{(m)}, x_{2j}^{(m)}, \ldots, x_{nj}^{(m)})'$, which is the $(j+1)^{th}$ column vector of the design matrix $X^{(m)}$, for $j = 1, 2, \ldots, k_m - 1$. Write $C_{0m} = C_{0m}(\tilde{\mathbf{x}}_1^{(m)}, \tilde{\mathbf{x}}_2^{(m)}, \ldots, \tilde{\mathbf{x}}_{k_m-1}^{(m)})$ and $C_m = C_m(\tilde{\mathbf{x}}_1^{(m)}, \tilde{\mathbf{x}}_2^{(m)}, \ldots, \tilde{\mathbf{x}}_{k_m-1}^{(m)})$.

Theorem 5: *The prior and posterior normalizing constants C_{0m} and C_m are scale-invariant in the covariates. Specifically, we have*

$$C_{0m}(\tilde{\mathbf{x}}_1^{(m)}, \tilde{\mathbf{x}}_2^{(m)}, \ldots, \tilde{\mathbf{x}}_{k_m-1}^{(m)}) = C_{0m}(a_1\tilde{\mathbf{x}}_1^{(m)}, a_2\tilde{\mathbf{x}}_2^{(m)}, \ldots, a_{k_m}\tilde{\mathbf{x}}_{k_m-1}^{(m)})$$

and

$$C_m(\tilde{\mathbf{x}}_1^{(m)}, \tilde{\mathbf{x}}_2^{(m)}, \ldots, \tilde{\mathbf{x}}_{k_m-1}^{(m)}) = C_m(a_1\tilde{\mathbf{x}}_1^{(m)}, a_2\tilde{\mathbf{x}}_2^{(m)}, \ldots, a_{k_m-1}\tilde{\mathbf{x}}_{k_m-1}^{(m)})$$

for all $a_1 > 0, a_2 > 0, \ldots, a_{k_m-1} > 0$.

Prior and Posterior Normalizing Constants

- For the logistic regression model, the prior normalizing constant is given by

$$C_0 = \int_{R^{k+1}} \left| X' W(\boldsymbol{\beta}) X \right|^{1/2} d\boldsymbol{\beta},$$

where $W(\boldsymbol{\beta}) = \text{diag}(w_1(\boldsymbol{\beta}), w_2(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta}))$, and $w_i(\boldsymbol{\beta}) = n_i \exp(\mathbf{x}_i' \boldsymbol{\beta}) / \{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})\}^2$.

Prior and Posterior Normalizing Constants

- For the logistic regression model, the prior normalizing constant is given by

$$C_0 = \int_{R^{k+1}} \left|X'W(\boldsymbol{\beta})X\right|^{1/2} d\boldsymbol{\beta},$$

where $W(\boldsymbol{\beta}) = \text{diag}(w_1(\boldsymbol{\beta}), w_2(\boldsymbol{\beta}), \ldots, w_n(\boldsymbol{\beta}))$, and $w_i(\boldsymbol{\beta}) = n_i \exp(\mathbf{x}_i'\boldsymbol{\beta})/\{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})\}^2$.

- The posterior normalizing constant can be written as

$$C = \int_{R^{k+1}} L(\boldsymbol{\beta}|X, \mathbf{y}) \left|X'W(\boldsymbol{\beta})X\right|^{1/2} d\boldsymbol{\beta},$$

where $L(\boldsymbol{\beta}|X, \mathbf{y}) = \prod_{i=1}^{n} \binom{n_i}{y_i} [\exp(y_i \mathbf{x}_i'\boldsymbol{\beta})/\{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})\}^{n_i}]$.

Logistic Regression Models with Binary Covariates

- We consider a saturated logistic regression model with s main binary covariates $x_{i1}, x_{i2}, \ldots, x_{is}$, each of which takes values of 0 or 1.

Logistic Regression Models with Binary Covariates

- We consider a saturated logistic regression model with s main binary covariates $x_{i1}, x_{i2}, \ldots, x_{is}$, each of which takes values of 0 or 1.
- We assume that in addition to an intercept and s main binary covariates, the model includes all possible interactions: $x_{ij}x_{ij'}$ $(j < j')$, $x_{ij}x_{ij'}x_{ij''}$ $(j < j' < j'')$, $\ldots$, $x_{i1}x_{i2}\ldots x_{is}$.

Logistic Regression Models with Binary Covariates

- We consider a saturated logistic regression model with s main binary covariates $x_{i1}, x_{i2}, \ldots, x_{is}$, each of which takes values of 0 or 1.
- We assume that in addition to an intercept and s main binary covariates, the model includes all possible interactions: $x_{ij}x_{ij'}$ $(j < j')$, $x_{ij}x_{ij'}x_{ij''}$ $(j < j' < j'')$, $\ldots$, $x_{i1}x_{i2}\ldots x_{is}$.
- In this case, $k = 2^s - 1$ and the total number of parameters including the intercept is 2^s.

Logistic Regression Models with Binary Covariates

For notational simplicity, we write

$$p_{x_1x_2\ldots x_s}(\boldsymbol{\beta}) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{s}\beta_j x_j + \sum_{j<j'} x_j x_{j'}\beta_{jj'} + \sum_{j<j'<j''} x_j x_{j'}\beta_{jj'} + + \cdots + x_1x_2\ldots x_s\beta_{12\cdots s}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{s}\beta_j x_j + \sum_{j<j'} x_j x_{j'}\beta_{jj'} + \sum_{j<j'<j''} x_j x_{j'} x_{j''}\beta_{jj'j''} + \cdots + x_1x_2\ldots x_s\beta_{12\cdots s}\right)},$$

where x_j takes the values 0 or 1 for $j = 1, 2, \ldots, s$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_s, \beta_{jj'}, 1 \le j < j' \le s, \ \ldots, \beta_{12\ldots s})'$. Then, we have $w_i(\boldsymbol{\beta}) = n_i p_{x_{i1}x_{i2}\ldots x_{is}}(\boldsymbol{\beta})\{1 - p_{x_{i1}x_{i2}\ldots x_{is}}(\boldsymbol{\beta})\}$.

Jeffreys's Prior with Binary Covariates

- Let $n_{(j_1 j_2 \ldots j_s)} = \sum_{i=1}^{n} n_i 1\{x_{i1} = j_1, x_{i2} = j_2, \ldots, x_{is} = j_s\}$ for $j_l = 0, 1$ and $l = 1, 2, \ldots, s$

Jeffreys's Prior with Binary Covariates

- Let $n_{(j_1 j_2 \ldots j_s)} = \sum_{i=1}^{n} n_i 1\{x_{i1} = j_1, x_{i2} = j_2, \ldots, x_{is} = j_s\}$ for $j_l = 0, 1$ and $l = 1, 2, \ldots, s$
- **Theorem 6**: *Under the saturated logistic regression model, Jeffreys's prior is proper if and only if $n_{(j_1 j_2 \ldots j_s)} \geq 1$ for all $j_l = 0, 1$, $l = 1, 2, \ldots, s$ and the kernel of the Jeffreys's prior in (1) reduces to*

$$|X'W(\beta)X|^{1/2} = \left(\prod_{j_1=0}^{1} \prod_{j_2=0}^{1} \cdots \prod_{j_s=0}^{1} \left[n_{(j_1 j_2 \ldots j_s)} p_{j_1 j_2 \ldots j_s}(\beta)\{1 - p_{j_1 j_2 \ldots j_s}(\beta)\} \right] \right)^{1/2}. \qquad (3)$$

Prior and Posterior Normalizing Constants

The normalizing constant for Jeffreys's prior has a closed form expression given by

$$C_0 = \left[\prod_{j_1=0}^{1}\prod_{j_2=0}^{1}\cdots\prod_{j_s=0}^{1} n_{(j_1 j_2 \ldots j_s)}\right]^{1/2}\left[B\left(\frac{1}{2},\frac{1}{2}\right)\right]^{2^s} = \pi^{2^s}\left[\prod_{j_1=0}^{1}\prod_{j_2=0}^{1}\cdots\prod_{j_s=0}^{1} n_{(j_1 j_2 \ldots j_s)}\right]^{1/2}.$$

The posterior normalizing constant based on Jeffreys's prior also has a closed form given as follows:

$$C = \int_{R^2} L(\boldsymbol{\beta}|X,\mathbf{y})|X'W(\boldsymbol{\beta})X|^{1/2}d\boldsymbol{\beta} = \left[\prod_{i=1}^{n}\binom{n_i}{y_i}\right]\left[\prod_{j_1=0}^{1}\prod_{j_2=0}^{1}\cdots\prod_{j_s=0}^{1} n_{(j_1 j_2 \ldots j_s)}\right]^{1/2}$$
$$\times\left\{\prod_{j_1=0}^{1}\prod_{j_2=0}^{1}\cdots\prod_{j_s=0}^{1} B\left[\frac{1}{2}+n^y_{(j_1 j_2 \ldots j_s)}, \frac{1}{2}+n_{(j_1 j_2 \ldots j_s)}-n^y_{(j_1 j_2 \ldots j_s)}\right]\right\},$$

where $n^y_{(j_1 j_2 \ldots j_s)} = \sum_{i=1}^{n} y_i 1\{x_{i1}=j_1, x_{i2}=j_2, \ldots, x_{is}=j_s\}$.

Connection between BIC and Normalizing Constants

Let $N = \sum_{i=1}^{n} n_i$ and

$$\hat{\alpha}_{(j_1 j_2 \dots j_s)} = \frac{n_{(j_1 j_2 \dots j_s)}}{N} \text{ and } \hat{\mu}_{(j_1 j_2 \dots j_s)} = \frac{n^y_{(j_1 j_2 \dots j_s)}}{N}.$$

for $j_l = 0, 1,\ l = 1, 2, \dots, s$. Also, let $\hat{\beta}$ denote the maximum likelihood estimate of β.
BIC is given by

$$\begin{aligned}
\text{BIC} = & -2 \log L(\hat{\beta}|X, \mathbf{y}) + 2^s \log(N) \\
= & -2 \log \left[\prod_{i=1}^{n} \binom{n_i}{y_i} \right] - 2 \sum_{j_1}^{1} \sum_{j_2=0}^{1} \cdots \sum_{j_s=0}^{1} \Bigg[n^y_{(j_1 j_2 \dots j_s)} \log \left\{ \frac{n^y_{(j_1 j_2 \dots j_s)}}{n_{(j_1 j_2 \dots j_s)}} \right\} \\
& + \{ n_{(j_1 j_2 \dots j_s)} - n^y_{(j_1 j_2 \dots j_s)} \} \log \left\{ \frac{n_{(j_1 j_2 \dots j_s)} - n^y_{(j_1 j_2 \dots j_s)}}{n_{(j_1 j_2 \dots j_s)}} \right\} \Bigg] + 2^s \log(N).
\end{aligned}$$

Connection between BIC and Normalizing Constants

- **Theorem 7**: *Assume that (i)* $\lim_{N\to\infty} \hat{\alpha}_{(j_1 j_2 \ldots j_s)} = \alpha_{(j_1 j_2 \ldots j_s)}$ *and* $\lim_{N\to\infty} \hat{\mu}_{(j_1 j_2 \ldots j_s)} = \mu_{(j_1 j_2 \ldots j_s)}$ *exist and (ii)* $0 < \alpha_{(j_1 j_2 \ldots j_s)} < 1$ *and* $0 < \mu_{(j_1 j_2 \ldots j_s)} < \alpha_{(j_1 j_2 \ldots j_s)}$ *for all* $j_l = 0, 1,\ l = 1, 2, \ldots, s$. *Then, for large* N*, we have*

$$-2(\log C - \log C_0) = \mathrm{BIC} + \sum_{j_1}^{1} \sum_{j_2=0}^{1} \cdots \sum_{j_s=0}^{1} \log\left[\frac{\pi}{2}\hat{\alpha}_{(j_1 j_2 \ldots j_s)}\right] + o\left(\frac{1}{N}\right).$$

Connection between BIC and Normalizing Constants

- **Theorem 7**: *Assume that (i)* $\lim_{N\to\infty} \hat{\alpha}_{(j_1 j_2 \ldots j_s)} = \alpha_{(j_1 j_2 \ldots j_s)}$ *and* $\lim_{N\to\infty} \hat{\mu}_{(j_1 j_2 \ldots j_s)} = \mu_{(j_1 j_2 \ldots j_s)}$ *exist and (ii)* $0 < \alpha_{(j_1 j_2 \ldots j_s)} < 1$ *and* $0 < \mu_{(j_1 j_2 \ldots j_s)} < \alpha_{(j_1 j_2 \ldots j_s)}$ *for all* $j_l = 0, 1,\ l = 1, 2, \ldots, s$. *Then, for large* N*, we have*

$$-2(\log C - \log C_0) = \mathrm{BIC} + \sum_{j_1}^{1} \sum_{j_2=0}^{1} \cdots \sum_{j_s=0}^{1} \log\left[\frac{\pi}{2}\hat{\alpha}_{(j_1 j_2 \ldots j_s)}\right] + o\left(\frac{1}{N}\right).$$

- $-2(\log C - \log C_0)$ acts very similarly to BIC.

Connection between BIC and Normalizing Constants

- **Theorem 7**: *Assume that (i)* $\lim_{N\to\infty}\hat{\alpha}_{(j_1 j_2 \ldots j_s)} = \alpha_{(j_1 j_2 \ldots j_s)}$ *and* $\lim_{N\to\infty}\hat{\mu}_{(j_1 j_2 \ldots j_s)} = \mu_{(j_1 j_2 \ldots j_s)}$ *exist and (ii)* $0 < \alpha_{(j_1 j_2 \ldots j_s)} < 1$ *and* $0 < \mu_{(j_1 j_2 \ldots j_s)} < \alpha_{(j_1 j_2 \ldots j_s)}$ *for all* $j_l = 0, 1$, $l = 1, 2, \ldots, s$. *Then, for large* N, *we have*

$$-2(\log C - \log C_0) = \mathrm{BIC} + \sum_{j_1}^{1} \sum_{j_2=0}^{1} \cdots \sum_{j_s=0}^{1} \log\left[\frac{\pi}{2}\hat{\alpha}_{(j_1 j_2 \ldots j_s)}\right] + o\left(\frac{1}{N}\right).$$

- $-2(\log C - \log C_0)$ acts very similarly to BIC.
- In addition to a dimensional penalty $2^s \log N$ in BIC, the dimensional penalty term in $-2(\log C - \log C_0)$ also depends on the "joint distribution" of covariates $(x_{i1}, x_{i2}, \ldots, x_{is})$.

Computation: Importance Sampling

- First we consider a more general form of the multivariate t-distribution with density

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

- For computing the prior normalizing constant, we specify $\mu = \mathbf{0}$ and match the curvatures of the Jeffreys's prior and the t-distribution at $\mathbf{0}$ as follows:

$$\kappa_0 \frac{\partial^2 \log \pi(\boldsymbol{\beta}|X)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\Big|_{\boldsymbol{\beta}=\mathbf{0}} = \frac{\partial^2 \log g(\boldsymbol{\beta}|\boldsymbol{\mu}=\mathbf{0},\Sigma,\nu)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\Big|_{\boldsymbol{\beta}=\mathbf{0}},$$

where $\kappa_0 > 0$ is a fixed scale-adjustment parameter.

Computation: Importance Sampling

- First we consider a more general form of the multivariate t-distribution with density

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

- For computing the prior normalizing constant, we specify $\mu = \mathbf{0}$ and match the curvatures of the Jeffreys's prior and the t-distribution at $\mathbf{0}$ as follows:

$$\kappa_0 \frac{\partial^2 \log \pi(\beta|X)}{\partial\beta\partial\beta'}\Big|_{\beta=\mathbf{0}} = \frac{\partial^2 \log g(\beta|\boldsymbol{\mu}=\mathbf{0},\Sigma,\nu)}{\partial\beta\partial\beta'}\Big|_{\beta=\mathbf{0}},$$

where $\kappa_0 > 0$ is a fixed scale-adjustment parameter.

Importance Sampling (continued)

- For computing the posterior normalizing constant, we specify

$$\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\beta} \in R^{k+1}}{\operatorname{argmax}} \{\log[L(\boldsymbol{\beta}|X, \mathbf{y})\pi(\boldsymbol{\beta}|X)]\}$$

and

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

where $\kappa_1 > 0$ is a fixed scale-adjustment parameter.

Importance Sampling (continued)

- To specify ν by matching a t-distribution to the square-root of the logistic distribution with a density that is proportional to $\sqrt{\exp(u)/\{1+\exp(u)\}^2}$. To do so, we match the curvatures at 0 and the percentiles of these two distributions, which gives $\nu = 3.37$.

Importance Sampling (continued)

- To specify ν by matching a t-distribution to the square-root of the logistic distribution with a density that is proportional to $\sqrt{\exp(u)/\{1+\exp(u)\}^2}$. To do so, we match the curvatures at 0 and the percentiles of these two distributions, which gives $\nu = 3.37$.
- We propose to use $\nu = 3.37$ as a guide value for ν in $g(\beta|\boldsymbol{\mu} = \mathbf{0}, \Sigma, \nu)$ for computing the prior normalizing constant.

Importance Sampling (continued)

- To specify ν by matching a t-distribution to the square-root of the logistic distribution with a density that is proportional to $\sqrt{\exp(u)/\{1+\exp(u)\}^2}$. To do so, we match the curvatures at 0 and the percentiles of these two distributions, which gives $\nu = 3.37$.
- We propose to use $\nu = 3.37$ as a guide value for ν in $g(\beta|\boldsymbol{\mu} = \mathbf{0}, \Sigma, \nu)$ for computing the prior normalizing constant.
- For the posterior normalizing constant, we specify $\nu \geq 3.37$ such as $\nu = 5$.

The Importance Sampling Algorithm for C_0

Step 1: Generate a random sample $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_Q)'$ of size Q from $g(\beta|\boldsymbol{\mu} = \mathbf{0}, \Sigma, \nu)$, where for each q, independently

(i) generate $\lambda_q \sim \mathcal{G}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$; and

(ii) generate $\boldsymbol{\beta}_q \sim N_{k+1}\left(\mathbf{0}, \Sigma/\lambda_q\right)$.

Step 2: Compute the Monte Carlo estimate of C_0 as

$$\hat{C}_0 = \frac{1}{Q} \sum_{q=1}^{Q} \frac{|X' W(\boldsymbol{\beta}_q) X|^{1/2}}{g(\boldsymbol{\beta}_q | \boldsymbol{\mu} = \mathbf{0}, \Sigma, \nu)}.$$

Comments

- In Step 2, we may also calculate $\log(\hat{C}_0)$ instead of $\hat{C}_0$.

Comments

- In Step 2, we may also calculate $\log(\hat{C}_0)$ instead of $\hat{C}_0$.
- In addition, we shall compute the relative MC standard error (RSE) as follows:

$$\text{RSE}(\hat{C}_0) = \frac{1}{\hat{C}_0}\left\{\frac{1}{Q(Q-1)}\sum_{q=1}^{Q}\left[\frac{|X'W(\boldsymbol{\beta}_q)X|^{1/2}}{g(\boldsymbol{\beta}_q|\boldsymbol{\mu}=\mathbf{0},\Sigma,\nu)} - \hat{C}_0\right]^2\right\}^{1/2}.$$

An Illustrative Example

We consider the logistic regression model with a binary covariate. We generate a simulated dataset of size $n = 100$. A summary of the simulated data is given as follows: $n_{(0)} = \sum_{i=1}^{n}(1 - x_{i1}) = 32$, $n_{(1)} = \sum_{i=1}^{n} x_{i1} = 68$, $\sum_{i=1}^{n}(1 - y_i) = 29$, $\sum_{i=1}^{n} y_i = 71$, $\sum_{i=1}^{n} y_i(1 - x_{i1}) = 19$, $\sum_{i=1}^{n}(n_i - y_i)(1 - x_{i1}) = 13$, $\sum_{i=1}^{n} y_i x_{i1} = 52$, and $\sum_{i=1}^{n}(n_i - y_i)x_{i1} = 16$. We implemented the proposed importance sampling algorithm with various values of κ_0 and $\kappa_1 5$. The results are given in Table 1.

Table 1. Monte Carlo estimates of $\log C_0$ and $\log C$

		Jeffreys's Prior			Posterior		
ν	MC Size (Q)	κ_0	$\log \hat{C}_0$	MC SE	κ_1	$\log \hat{C}$	MC SE
1	5,000	1	6.143	0.011	2	-56.905	0.009
	10,000		6.137	0.008		-56.907	0.007
3.37	5,000		6.130	0.003		-56.906	0.005
	10,000		6.131	0.002		-56.900	0.003
5	5,000		6.134	0.003		-56.896	0.004
	10,000		6.133	0.002		-56.895	0.003
10	5,000		6.140	0.008		-56.879	0.006
	10,000		6.143	0.006		-56.881	0.004
20	5,000		6.146	0.015		-56.877	0.009
	10,000		6.139	0.012		-56.881	0.007
3.37	5,000	0.5	6.145	0.008	1	-56.883	0.008
	10,000		6.144	0.005		-56.884	0.006
	5,000	2	6.135	0.008	3	-56.914	0.006
	10,000		6.127	0.005		-56.906	0.004
5	5,000	0.5	6.129	0.006	1	-56.901	0.006
	10,000		6.129	0.004		-56.898	0.004
	5,000	2	6.141	0.012	3	-56.889	0.007
	10,000		6.139	0.008		-56.890	0.005
true values		$\log C_0 = 6.132$			$\log C = -56.890$		

Figure 1

- Figure 1 shows the densities of the Jeffreys's prior and the corresponding posterior distribution.

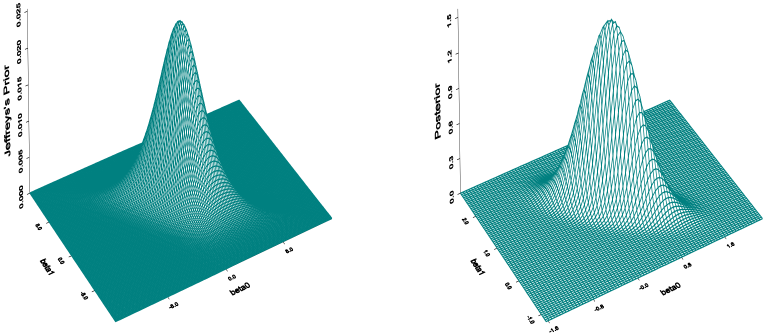
Figure 1

- Figure 1 shows the densities of the Jeffreys's prior and the corresponding posterior distribution.
- From Figure 1, we see that the Jeffreys's prior is unimodal and symmetric about 0.

Figure 1

- Figure 1 shows the densities of the Jeffreys's prior and the corresponding posterior distribution.
- From Figure 1, we see that the Jeffreys's prior is unimodal and symmetric about 0.
- The height of the Jeffreys's prior is quite small, indicating that the prior is quite flat.

The densities of the Jeffreys's prior (left) and the posterior distribution in (right).



Simulation Design

- For each simulated data set, n independent Bernoulli observations y_i's are generated with success probability

$$p_i = \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\}},$$

Simulation Design

- For each simulated data set, n independent Bernoulli observations y_i's are generated with success probability

$$p_i = \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\}},$$

- $(x_{1i}, x_{2i}, x_{3i}, x_{4i})'$ are *i.i.d.* random vectors such that $x_{1i} \sim \text{Ber}(p_{1i})$, $x_{2i}|x_{1i} \sim \text{Ber}(p_{2i})$, and $(x_{3i}, x_{4i})'|x_{1i}, x_{2i} \sim N\left\{ \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}.$

Simulation Design

- For each simulated data set, n independent Bernoulli observations y_i's are generated with success probability

$$p_i = \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}\}},$$

- $(x_{1i}, x_{2i}, x_{3i}, x_{4i})'$ are $i.i.d.$ random vectors such that $x_{1i} \sim \text{Ber}(p_{1i})$, $x_{2i}|x_{1i} \sim \text{Ber}(p_{2i})$, and $(x_{3i}, x_{4i})'|x_{1i}, x_{2i} \sim N\left\{ \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$.
- We take $p_{1i} = 0.5$, $p_{2i} = \frac{\exp(0.5+0.6x_{1i})}{1+\exp(0.5+0.6x_{1i})}$, $\mu_{1i} = 0.1x_{1i} + 0.2x_{2i}$, $\mu_{2i} = -0.2x_{1i} - 0.1x_{2i}$.

Two Simulations

- In Simulation I, we use $\rho = 0.8$, and $\boldsymbol{\beta} = (0.1, 0, 0.5, 0, 0)'$, $\boldsymbol{\beta} = (0.1, 0, 0.5, -1.0, 0)'$, $\boldsymbol{\beta} = (0.1, 0, 0.5, -1.0, 2.5)'$, and $\boldsymbol{\beta} = (0.1, 1.5, 0.5, -1.0, 2.5)'$, which correspond to the true models (x_2), (x_2, x_3), (x_2, x_3, x_4), and (x_1, x_2, x_3, x_4) (full model), respectively.

Two Simulations

- In Simulation I, we use $\rho = 0.8$, and $\boldsymbol{\beta} = (0.1, 0, 0.5, 0, 0)'$, $\boldsymbol{\beta} = (0.1, 0, 0.5, -1.0, 0)'$, $\boldsymbol{\beta} = (0.1, 0, 0.5, -1.0, 2.5)'$, and $\boldsymbol{\beta} = (0.1, 1.5, 0.5, -1.0, 2.5)'$, which correspond to the true models (x_2), (x_2, x_3), (x_2, x_3, x_4), and (x_1, x_2, x_3, x_4) (full model), respectively.
- In Simulation II, we use $\rho = 0.7$, and $\boldsymbol{\beta} = (1.0, 0, -1.3, 0, 0)'$, $\boldsymbol{\beta} = (1.0, 0, -1.3, 1.0, 0)'$, $\boldsymbol{\beta} = (1.0, 0, -1.3, 1.0, 1.7)'$, and $\boldsymbol{\beta} = (1.0, 1.5, -1.3, 1.0, 1.7)'$, which correspond to the true models (x_2), (x_2, x_3), (x_2, x_3, x_4), and (x_1, x_2, x_3, x_4) (full model), respectively.

Comments on Two Simulations

- The differences in the regression coefficients are greater in Simulation II than in Simulation I.

Comments on Two Simulations

- The differences in the regression coefficients are greater in Simulation II than in Simulation I.
- x_{i3} and x_{i4} are less correlated in Simulation II than in Simulation I.

Comments on Two Simulations

- The differences in the regression coefficients are greater in Simulation II than in Simulation I.
- x_{i3} and x_{i4} are less correlated in Simulation II than in Simulation I.
- We expect that the methods (criteria) should perform better in Simulation II than in Simulation I.

Other Details

- We use sample sizes of $n = 100$, $n = 250$, and $n = 500$.

Other Details

- We use sample sizes of $n = 100$, $n = 250$, and $n = 500$.
- Under each simulation design, for each combination of (n, β), we independently generate $N = 500$ datasets.

Other Details

- We use sample sizes of $n = 100$, $n = 250$, and $n = 500$.
- Under each simulation design, for each combination of (n, β), we independently generate $N = 500$ datasets.
- For each simulated dataset, we fit $2^4 = 16$ models.

Other Details

- We use sample sizes of $n = 100$, $n = 250$, and $n = 500$.
- Under each simulation design, for each combination of (n, β), we independently generate $N = 500$ datasets.
- For each simulated dataset, we fit $2^4 = 16$ models.
- The Monte Carlo sample of size is $Q = 20,000$.

Other Details

- We use sample sizes of $n = 100$, $n = 250$, and $n = 500$.
- Under each simulation design, for each combination of (n, β), we independently generate $N = 500$ datasets.
- For each simulated dataset, we fit $2^4 = 16$ models.
- The Monte Carlo sample of size is $Q = 20,000$.
- We compute posterior model probabilities under Jeffreys's prior and g-type prior (Zellner, 1986), which is defined as

$$\pi_g(\beta|\mathbf{X}) = \frac{|X'X|^{1/2}}{(2\pi\tau_0)^{(k+1)/2}} \exp\left\{-\frac{1}{2\tau_0}\beta'(X'X)\beta\right\}.$$

Other Details

- We use sample sizes of $n = 100$, $n = 250$, and $n = 500$.
- Under each simulation design, for each combination of (n, β), we independently generate $N = 500$ datasets.
- For each simulated dataset, we fit $2^4 = 16$ models.
- The Monte Carlo sample of size is $Q = 20,000$.
- We compute posterior model probabilities under Jeffreys's prior and g-type prior (Zellner, 1986), which is defined as

$$\pi_g(\beta|\mathbf{X}) = \frac{|X'X|^{1/2}}{(2\pi\tau_0)^{(k+1)/2}} \exp\Big\{ -\frac{1}{2\tau_0}\beta'(X'X)\beta \Big\}.$$

- We also compute AIC and BIC.

Table 2. Frequencies for Ranking the True Model as Best Based on $N = 500$ Datasets

		Simulation I			
n	True model	Jeffreys's Prior	g-Type Prior	AIC	BIC
100	(x_2)	110	231	118	76
	(x_2, x_3)	85	35	128	61
	(x_2, x_3, x_4)	47	7	110	33
	(x_1, x_2, x_3, x_4)	29	1	118	19
250	(x_2)	156	325	185	121
	(x_2, x_3)	133	74	189	105
	(x_2, x_3, x_4)	93	37	191	77
	(x_1, x_2, x_3, x_4)	95	26	258	66
500	(x_2)	295	416	291	261
	(x_2, x_3)	233	173	292	198
	(x_2, x_3, x_4)	179	127	304	163
	(x_1, x_2, x_3, x_4)	179	118	359	152

Table 2. Frequencies for Ranking the True Model as Best Based on $N = 500$ Datasets (continued)

		Simulation II			
n	True model	Jeffreys's Prior	g-Type Prior	AIC	BIC
100	(x_2)	363	461	274	355
	(x_2, x_3)	321	231	300	299
	(x_2, x_3, x_4)	179	59	244	141
	(x_1, x_2, x_3, x_4)	106	26	255	92
250	(x_2)	465	487	310	474
	(x_2, x_3)	463	464	353	469
	(x_2, x_3, x_4)	420	347	398	400
	(x_1, x_2, x_3, x_4)	388	274	481	362
500	(x_2)	472	490	304	487
	(x_2, x_3)	484	493	365	488
	(x_2, x_3, x_4)	487	485	421	486
	(x_1, x_2, x_3, x_4)	489	478	499	486

The Data

- Data are from a retrospective cohort study of men treated with radical prostatectomy ($n = 968$) between 1988-2000 (D'Amico et al., 2002).

The Data

- Data are from a retrospective cohort study of men treated with radical prostatectomy ($n = 968$) between 1988-2000 (D'Amico et al., 2002).
- The Binary response is PECE, which takes the values 0 and 1, where a 1 denotes that the cancer has penetrated the prostate wall and a 0 indicates otherwise.

The Data

- Data are from a retrospective cohort study of men treated with radical prostatectomy ($n = 968$) between 1988-2000 (D'Amico et al., 2002).
- The Binary response is PECE, which takes the values 0 and 1, where a 1 denotes that the cancer has penetrated the prostate wall and a 0 indicates otherwise.
- The covariates include age, Log(PSA), ppb (percent positive prostate biopsies), biopsy Gleason score (GG7, GG8H), and clinical tumor stage (T2b,T2c).

Variable Selection

- We compare 32 models.

Variable Selection

- We compare 32 models.
- The best model under BIC, and the model with the highest posterior probability based on both the Jeffreys's prior and the g-type prior is (LogPSA, ppb, GG7, GG8H).

Variable Selection

- We compare 32 models.
- The best model under BIC, and the model with the highest posterior probability based on both the Jeffreys's prior and the g-type prior is (LogPSA, ppb, GG7, GG8H).
- For this best model, the posterior probability is 0.806 and 0.828 for the Jeffreys's prior and the g-type prior, respectively.

Variable Selection

- We compare 32 models.
- The best model under BIC, and the model with the highest posterior probability based on both the Jeffreys's prior and the g-type prior is (LogPSA, ppb, GG7, GG8H).
- For this best model, the posterior probability is 0.806 and 0.828 for the Jeffreys's prior and the g-type prior, respectively.
- The AIC criterion selects the full model (age, LogPSA, ppb, GG7, GG8H, T2b, T2c) as the best model.

Computing HPD interval via Importance Sampling

- We use the Monte Carlo method proposed by Chen and Shao (1999).

Computing HPD interval via Importance Sampling

- We use the Monte Carlo method proposed by Chen and Shao (1999).
- Let $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_Q\}$, where $\boldsymbol{\beta}_q = (\beta_{q0}, \beta_{q1}, \ldots, \beta_{qk})'$, $q = 1, 2, \ldots, Q$, be a random sample of size Q from $g(\boldsymbol{\beta}|\boldsymbol{\mu}, \Sigma, \nu)$.

Computing HPD interval via Importance Sampling

- We use the Monte Carlo method proposed by Chen and Shao (1999).
- Let $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_Q\}$, where $\boldsymbol{\beta}_q = (\beta_{q0}, \beta_{q1}, \ldots, \beta_{qk})'$, $q = 1, 2, \ldots, Q$, be a random sample of size Q from $g(\boldsymbol{\beta}|\boldsymbol{\mu}, \Sigma, \nu)$.
- The posterior density is $\pi(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \propto \pi^*(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})|X'W(\boldsymbol{\beta})X|$.

Computing HPD interval via Importance Sampling

- We use the Monte Carlo method proposed by Chen and Shao (1999).
- Let $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_Q\}$, where $\boldsymbol{\beta}_q = (\beta_{q0}, \beta_{q1}, \ldots, \beta_{qk})'$, $q = 1, 2, \ldots, Q$, be a random sample of size Q from $g(\boldsymbol{\beta}|\boldsymbol{\mu}, \Sigma, \nu)$.
- The posterior density is $\pi(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \propto \pi^*(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})|X'W(\boldsymbol{\beta})X|$.
- Let $\omega_q = \frac{\pi^*(\boldsymbol{\beta}_q|\mathbf{X},\mathbf{y})}{g(\boldsymbol{\beta}_q|\boldsymbol{\mu},\Sigma,\nu)}$ for $q = 1, 2, \ldots, Q$.

Computing HPD interval via Importance Sampling

- We use the Monte Carlo method proposed by Chen and Shao (1999).
- Let $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_Q\}$, where $\boldsymbol{\beta}_q = (\beta_{q0}, \beta_{q1}, \ldots, \beta_{qk})'$, $q = 1, 2, \ldots, Q$, be a random sample of size Q from $g(\boldsymbol{\beta}|\boldsymbol{\mu}, \Sigma, \nu)$.
- The posterior density is $\pi(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \propto \pi^*(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})|X'W(\boldsymbol{\beta})X|$.
- Let $\omega_q = \frac{\pi^*(\boldsymbol{\beta}_q|\mathbf{X},\mathbf{y})}{g(\boldsymbol{\beta}_q|\boldsymbol{\mu},\Sigma,\nu)}$ for $q = 1, 2, \ldots, Q$.
- We compute the highest posterior density (HPD) interval of β_j.

The Algorithm

- For $0 \leq \gamma < 1$, define

$$\hat{\beta}_j^{(\gamma)} = \begin{cases} \beta_{j(1)} & \text{if } \gamma = 0, \\ \beta_{j(q)} & \text{if } \sum_l^{q-1} \omega_l < \gamma \leq \sum_{l=1}^{q} \omega_l, \end{cases}$$

where $\beta_{j(q)}$ is the q^{th} smallest of $\{\beta_{j(l)},\ l = 1, 2, \ldots, Q\}$.

The Algorithm

- For $0 \leq \gamma < 1$, define

$$\hat{\beta}_j^{(\gamma)} = \begin{cases} \beta_{j(1)} & \text{if } \gamma = 0, \\ \beta_{j(q)} & \text{if } \sum_l^{q-1} \omega_l < \gamma \leq \sum_{l=1}^{q} \omega_l, \end{cases}$$

  where $\beta_{j(q)}$ is the q^{th} smallest of $\{\beta_{j(l)},\ l = 1, 2, \ldots, Q\}$.
- To obtain a $100(1-\alpha)\%$ HPD interval for β_j, we let

$$R_q(Q) = \left(\hat{\beta}_j^{(\frac{q}{Q})}, \hat{\beta}_j^{(\frac{q+[(1-\alpha)Q]}{Q})} \right)$$

  for $q = 1, 2, \ldots, Q - [(1-\alpha)Q]$, where $[(1-\alpha)Q]$ denotes the integer part of $(1-\alpha)Q$.

The Algorithm

- For $0 \leq \gamma < 1$, define

$$\hat{\beta}_j^{(\gamma)} = \begin{cases} \beta_{j(1)} & \text{if } \gamma = 0, \\ \beta_{j(q)} & \text{if } \sum_l^{q-1} \omega_l < \gamma \leq \sum_{l=1}^{q} \omega_l, \end{cases}$$

where $\beta_{j(q)}$ is the q^{th} smallest of $\{\beta_{j(l)},\ l = 1, 2, \ldots, Q\}$.

- To obtain a $100(1-\alpha)\%$ HPD interval for β_j, we let

$$R_q(Q) = \left(\hat{\beta}_j^{\left(\frac{q}{Q}\right)}, \hat{\beta}_j^{\left(\frac{q+[(1-\alpha)Q]}{Q}\right)}\right)$$

for $q = 1, 2, \ldots, Q - [(1-\alpha)Q]$, where $[(1-\alpha)Q]$ denotes the integer part of $(1-\alpha)Q$.

- Then, the $100(1-\alpha)\%$ HPD interval is $R_{q^*}(Q)$, which is the interval that has the smallest width among all $R_q(Q)$'s.

Table 3. Estimates of the β_j's under Model (LogPSA, ppb, GG7, GG8H)

Variable	Maximum Likelihood Estimates			Posterior Estimates		
	Estimate	SE	p-value	Estimate	SE	95% HPD Interval
Intercept	-3.895	0.304	$<$0.0001	-3.896	0.307	(-4.586, -3.222)
LogPSA	0.696	0.135	$<$0.0001	0.696	0.135	(0.400, 1.004)
ppb	2.376	0.355	$<$0.0001	2.376	0.356	(1.612, 3.201)
G7	0.705	0.182	0.0001	0.706	0.182	(0.283, 1.098)
G8H	1.420	0.337	$<$0.0001	1.420	0.337	(0.639, 2.156)

Concluding Remarks

- We have undertaken a detailed theoretical investigation of Jeffreys's prior and have demonstrated its properties and performance in variable selection.

Concluding Remarks

- We have undertaken a detailed theoretical investigation of Jeffreys's prior and have demonstrated its properties and performance in variable selection.
- The prior has tails that are always in between multivariate t and multivariate normal distributions under logistic regression, regardless of the sample size or the dimension of $\boldsymbol{\beta}$.

Concluding Remarks

- We have undertaken a detailed theoretical investigation of Jeffreys's prior and have demonstrated its properties and performance in variable selection.
- The prior has tails that are always in between multivariate t and multivariate normal distributions under logistic regression, regardless of the sample size or the dimension of $\boldsymbol{\beta}$.
- The prior and posterior normalizing constants are scale invariant with respect to the covariates.

Concluding Remarks

- We have undertaken a detailed theoretical investigation of Jeffreys's prior and have demonstrated its properties and performance in variable selection.
- The prior has tails that are always in between multivariate t and multivariate normal distributions under logistic regression, regardless of the sample size or the dimension of $\boldsymbol{\beta}$.
- The prior and posterior normalizing constants are scale invariant with respect to the covariates.
- The prior only requires importance sampling to get accurate estimates of posterior model probabilities.

Thank You!