The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 523-532

Bayesian Methods for Generalized Linear Models

Paul E. Green[1], Dae Hak Kim[2], Dae Kee Min[3]

Abstract

Generalized linear models have various applications for data arising from many kinds of statistical studies. Although the response variable is generally assumed to be generated from a wide class of probability distributions, we focus on count data that are most often analyzed using binomial models for proportions or poisson models for rates. The methods and results presented here also apply to many other categorical data models in general, due to the relationship between multinomial and poisson sampling. The novelty of the approach suggested here is that all conditional distributions can be specified directly so that straightforward Gibbs sampling is possible. The prior distribution consists of two stages. We rely on a normal nonconjugate prior at the first stage, and a vague prior for hyperparameters at the second stage. The methods are demonstrated with an illustrative example using data collected by Rosenkranz and Raftery(1994) concerning the number of hospital admissions due to back pain in Washington state.

1. Introduction

The major impediment to the routine Bayesian implementation of generalized linear models has been the calculation of high dimensional integrals. Before the advent of sampling based methods for calculating marginal densities, possible solutions included numerical integration techniques, such as those introduced by Naylor and Smith(1982), or analytic approximations and their improvements suggested by Tierney and Kadane(1986) and Tierney, Kass, and Kadane(1989). In addition to being quite accurate, application of analytic approximations based on Laplace's method often provides insight into asymptotic properties of posterior inference. In order to use Laplace's method however, the log posterior must be dominated by a single mode, and a second order Taylor series expansion must be evaluated at the maximum. Thus, a difficult integral calculation is replaced by a much easier maximization. However, for high dimensional problems even a maximization can be cumbersome.

With the introduction of sampling based methods such as those proposed by Gelfand and

1) Assistant Professor, Department of Statistical Informtion, Catholic University of Taegu-Hyosung, Kyungsan, 712-713, Korea
2) Associate Professor, Department of Statistical Informtion, Catholic University of Taegu-Hyosung, Kyungsan, 712-713, Korea
3) Full time lecturer, Department of Statistical Informtion, Catholic University of Taegu-Hyosung, Kyungsan, 712-713, Korea

Smith(1990), Bayesian inference procedures changed dramatically. The idea is to simulate a Markov chain that converges to a target or stationary distribution, which in Bayesian data analysis is most often a posterior or predictive distribution. After the chain converges, a sample, perhaps correlated, is essentially being drawn from the target distribution. Any characteristics of the distribution can then be derived directly from the sample. Since the introduction of the work by Gelfand and Smith, sampling based methods, now commonly referred to as Markov Chain Monte Carlo(MCMC), have been applied to a wide variety of statistical models with great success. Today, MCMC and its variants continue to be the focus of much research activity.

The choice of a particular sampler depends on the problem under consideration. A sampler that mixes well in one case may perform poorly in another. For each sampler there are advantages and disadvantages. The amount of literature related to MCMC is huge, and a complete list of references is too large to enumerate here, but a good starting point includes the work by Gelman *et al.*(1995), and Gilks *et al.*(1996). The two samplers most commonly used today are the Metropolis and Gibbs samplers. The Metropolis sampler requires specification of a candidate distribution, which in some sense should be close to the target. Random draws are generated from the candidate and are accepted or rejected according to the calculation of a jump probability. The Gibbs sampler, a special case of Metropolis, accepts new candidates with probability one and is easy to implement, but requires the ability to sample directly from various conditional distributions. Other samplers in common use include the independence sampler(Tierney, 1994) and the hit-and-run sampler(Schmeiser and Chen, 1991).

Output from MCMC simulation should be monitored carefully. The model or the sampler may require fine tuning in order to accelerate convergence, promote rapid mixing, or in the case of Metropolis, improve acceptance rates. Model checking and ensuring propriety of posterior distributions are important aspects of any Bayesian data analysis. A reparameterization can also be useful for improving the efficiency of an MCMC sampler. Thus, the trial and error nature of sampling based methods for Bayesian data modeling makes software development a challenging task. One ambitious endeavor in this regard is BUGS (Bayesian inference Using Gibbs Sampling) software developed by Spiegelhalter *et al.*(1996).

Bayesian calculations are usually more tractible when using conjugate priors. For binomial and poisson likelihoods this requires consideration of the beta and gamma densities, respectively. However, when making posterior and predictive inference, the hyperparameters are embedded in gamma functions. While this is not entirely problematic, it does restrict the choice of sampling methods available to the data analyst. By using a nonconjugate normal prior at the first stage, and modifying the likelihood to match a normal distribution on a suitably transformed scale, Bayesian data analysis of a generalized linear model becomes equivalent to Bayesian data analysis of a normal linear model. At that point, all results of the normal model become available. In addition, all conditional distributions can be specified and sampled from directly so that straightforward Gibbs sampling is possible.

Two issues concerning the methods presented here deserve explanation. First, modification of the likelihood does not compromise propriety of the posterior since it coincides with a normal likelihood on a transformed scale. Second, without loss of generality, we choose to utilize the usual noninformative prior for hyperparameters commonly used in regression models, assuming that the sample size is large relative to the number of parameters. For those who prefer proper priors, or are attempting Bayesian analysis of more ambitious models such as neural networks or mixture models, the extension to proper priors poses no difficulty and can proceed according to the application.

In Section 2 the generalized linear model is reviewed. The Bayesian model is presented in Section 3 and an approximation to the likelihood and the conditional distributions needed to implement the Gibbs sampler are described. In Section 4 a data example is presented and Section 5 concludes with a discussion.

2. Generalized Linear Models

We assume the reader is familiar with generalized linear models, hereafter referred to as GLMs. This section will be brief and is presented to make terminology and notation precise in the sections that follow. Good accounts of GLMs appear in Agresti(1990) and McCullagh and Nelder(1989).

In the GLM framework, independent observations are available on the response variable $\mathbf{y}^T = (y_1, \cdots, y_N)$. The observations are assumed to be generated from an exponential family of the form

$$p(y_i \mid \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\} \tag{2.1}$$

where θ_i is called the *natural parameter* and ϕ is a scale parameter.

As in the normal linear model, the right side of a GLM is a linear function of explanatory variables $\mathbf{x}_i^T = (x_{i1}, \cdots, x_{ip})$ and parameters $\boldsymbol{\beta}^T = (\beta_1, \cdots, \beta_p)$ known as the *linear predictor*

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}. \tag{2.2}$$

A monotonic and differentiable link function g relates the expectation μ_i of y_i to the linear predictor via

$$g(\mu_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta} \tag{2.3}$$

When $\theta_i = \eta_i$ the function g is called the *canonical link* function. In the sections that follow we will denote the contribution of a single observation to the likelihood by $p = (y_i \mid \eta_i)$, since

η_i is related to θ_i through g. Dependence on ϕ is suppressed since a scale parameter will be introduced in the prior distribution of the Bayesian model in the next section.

In general, estimation for GLMs requires iterative methods and maximum likelihood estimates of β can be calculated using *adjusted dependent variable regression*, a form of iteratively reweighted least squares(IRLS). Model checking can be assessed with a likelihood ratio statistic called the *deviance* or Pearson's chi-square statistic. Outliers can be detected using the individual components of the deviance or chi-square statistics.

3. The Bayesian Model

A hierarchical model with a two stage prior distribution is described for Bayesian data analysis of GLMs. At the first stage, the linear predictor η_i is assigned a normal distribution in order to be compatible with the normal linear model. The second stage prior is given the usual noninformative prior. Formally, the model is

$$y_i \mid \eta_i \sim \text{(any exponential family)} \quad i=1,\cdots,N$$

$$\eta_i \mid \beta, \sigma^2 \sim N(\lambda_i, \sigma^2)$$

$$\eta_i = g(\mu_i) = \lambda_i + \varepsilon_i, \quad \varepsilon_i \sim N(0,\sigma^2), \quad \lambda_i = \mathbf{x}_i{}^T \beta$$

$$p(\beta, \sigma^2) \propto (\sigma^2)^{-1}.$$

By any exponential family we mean normal, binomial, poisson, or gamma response variables. The methods are not applicable, for example, for binary or multinomial ordered response models (see, for example, Albert and Chib(1993) for Bayesian analysis of these models). If a likelihood already contains a scale parameter, such as the normal or gamma, it should be set to a constant since σ^2 is included in the prior. Without loss of generality, this constant can be set to one. It is understood that $\phi=1$ for binomial and poisson likelihoods.

Under these assumptions, the joint posterior distribution is

$$p(\eta, \beta, \sigma^2 \mid y) \propto (\sigma^2)^{-(N/2+1)} [\prod p(y_i \mid \eta_i)] \exp\left[-\frac{1}{2\sigma^2}\sum(\eta_i-\lambda_i)^2\right]. \tag{3.1}$$

As in Bayesian analysis of the normal linear model, the following results are immediate

$$\beta \mid \eta, \sigma^2, y \sim N_p(\hat{\beta}, \sigma^2(X^TX)^{-1}), \quad \hat{\beta}=(X^TX)^{-1}X^T\eta$$

$$\sigma^2 \mid \eta, y \sim \text{Inv-}\chi^2(N-p, s^2), \quad s^2=\frac{1}{N-p}(\eta-X\hat{\beta})^T(\eta-X\hat{\beta})$$

where X is the $N\times p$ design matrix with rows x_i, and $\text{Inv}-\chi^2(\nu, s^2)$ denotes the scaled inverse-chi-square distribution with scale s^2 and degrees of freedom ν. Thus, we can sample directly from two of the three distributions necessary to implement Gibbs sampling. The next subsection describes a method to sample from $\eta \mid \beta, \sigma^2, y$ which completes the specification.

3.1 Approximating the Likelihood

The likelihood will now be matched to a normal distribution with respect to η_i on the scale of the link function. Let $p(\eta_i)$ denote the log of $p(y_i \mid \eta_i)$ and consider a second order Taylor series expansion about the value $\hat{\eta}_i$ that maximizes $p(\eta_i)$. Then

$$p(\eta_i) \approx p(\hat{\eta}_i) - \frac{1}{2} I(\hat{\eta}_i)(\eta_i - \hat{\eta}_i)^2$$

where

$$I(\hat{\eta}_i) = -\left. \frac{\partial^2 p(\eta_i)}{\partial \eta_i^2} \right|_{\eta_i = \hat{\eta}_i}$$

denotes the observed information evaluated at the maximum. Now, replace $p(y_i \mid \eta_i)$ with $\exp[p(\eta_i)]$ and the joint posterior distribution can be written as

$$p(\eta, \beta, \sigma^2 \mid y) \propto (\sigma^2)^{-(N/2+1)} \exp\left[-\frac{1}{2} \left(\sum_{i=1}^{N} \frac{(\eta_i - \hat{\eta}_i)^2}{[I(\hat{\eta}_i)]^{-1}} + \frac{(\eta_i - \lambda_i)^2}{\sigma^2} \right) \right]. \tag{3.2}$$

After collecting terms and completing the square

$$\eta_i \mid \beta, \sigma^2, y \sim N(\theta_i, \delta_i^2) \tag{3.3}$$

where

$$\frac{\theta_i = \frac{1}{\sigma^2}\lambda_i + I(\hat{\eta}_i)\hat{\eta}_i}{\frac{1}{\sigma^2} + I(\hat{\eta}_i)}, \qquad \delta_i^2 = \left(\frac{1}{\sigma^2} + I(\hat{\eta}_i) \right)^{-1}.$$

Note that the mean θ_i is a weighted average of the linear predictor λ_i in the prior distribution and the sample data $\hat{\eta}_i$. The maximum $\hat{\eta}_i$ is simply the link function evaluated at the data. For example, if $y_i \mid \eta_i \sim$ binomial (n_i, p_i) and logit $(p_i) = \eta_i$, then $\hat{\eta}_i$ is the sample logit. If $y_i = 0$ or $n_i = y_i$, the empirical logit can be used

$$\hat{\eta}_i = \log\left(\frac{y_i + 1/2}{n_i - y_i + 1/2} \right). \tag{3.4}$$

If $y_i \mid \eta_i \sim$ poisson (μ_i) and $\log \mu_i = \eta_i$, then $\hat{\eta}_i = \log y_i$. If $y_i = 0$ then

$$\hat{\eta}_i = \log(y_i + 1/2) \tag{3.5}$$

can be used. Thus, in some cases it may be necessary to make small adjustments to the data to ensure that the link function is defined at all values. This presents no difficulty since a small modification can actually improve the approximation. This assertion will be made clear in the data example.

3.2 Using the Gibbs Sampler

Using the approximation presented above, all conditionals are available as standard distributions and the following sampling scheme is presented to implement the Gibbs sampler.

Specify starting values β_0, σ_0^2 ;

Draw $\eta_i \mid \beta, \sigma^2, y \sim N(\theta_i, \delta_i^2)$

Draw $\sigma^2 \mid \eta, y \sim \text{Inv}-\chi^2(N-p, s^2)$

Draw $\beta \mid \eta, \sigma^2, y \sim N_p(\hat{\beta}, \sigma^2(X^TX)^{-1})$

Repeat for the desired number of iterations.

Experience has shown this sampler to perform well for various data problems. Very little burn-in is required even when poor starting values are chosen for β and σ^2.

4. Data Example: A Poisson Model for Rates

Rosenkranz and Raftery(1994) provide data concerning the number of hospital admissions for the medical treatment of back pain by county size (exposure) for the thirty-nine counties of Washington state during calendar year 1989. A question of interest is whether hospital admission rates are more related to availability of medical resources or to actual need. Rosenkranz and Raftery showed that admission rates were more related to number of hospitals per 10,000 residents than to the proportion of the population sixty-five years of age or older, suggesting that rates are more related to availability of medical resources. We use these data to fit a poisson model with a log link, using number of hospitals per 10,000 residents as a covariate.

Let y_i denote the number of admissions due to back pain for county i. Let t_i(exposure) be the number of adult residents ($\geq$ 20 years of age), and x_i the number of hospitals per 10,000 residents. Under the framework established in Section 3

$$y_i \mid t_i, \eta_i \sim \text{Poisson}(t_i e^{\eta_i})$$
$$\eta_i \mid \beta_0, \beta_1, \sigma^2 \sim N(\lambda_i, \sigma^2), \quad \lambda_i = \beta_0 + \beta_1 x_i$$
$$p(\beta_0, \beta_1, \sigma^2) \propto (\sigma^2)^{-1}.$$

In order to compare the normal approximation to the poisson likelihood, the exact joint posterior distribution is

$$p(\boldsymbol{\eta}, \beta_0, \beta_1, \sigma^2 \mid y) \propto (\sigma^2)^{-(N/2+1)} \exp\left[\sum_{i=1}^{N} y_i \eta_i - t_i e^{\eta_i} - \frac{1}{2\sigma^2} (\eta_i - \lambda_i)^2 \right] \tag{4.1}$$

Note that the contribution from the likelihood with respect to η_i resembles an extreme value distribution(if $y_i = t_i = 1$ the likelihood is a standard extreme value distribution). An overlay is shown in Figure 1 that compares the exact likelihood with the normal approximation when $y_i = 5$ and $t_i = 10$. The exact likelihood, sharing properties with the extreme value distribution, is skewed slightly to the left. As asserted in Section 3, a small modification to the data can improve the normal approximation by shifting the mean to the left and increasing the variance.

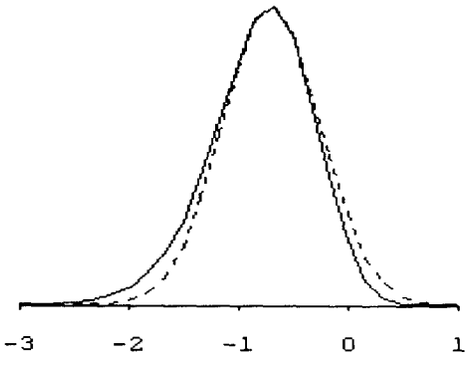

Figure 1. Overlay of normal approximation (dotted line) and exact likelihood

After making the approximation, the joint posterior distribution is

$$p(\boldsymbol{\eta}, \beta_0, \beta_1, \sigma^2 \mid y) \propto (\sigma^2)^{-(N/2+1)} \exp\left[-\frac{1}{2} \left(\sum_{i=1}^{N} \frac{(\eta_i - \log(y_i/t_i))^2}{1/y_i} + \frac{(\eta_i - \lambda_i)^2}{\sigma^2} \right) \right].$$

For the hospital data all $y_i > 0$ and an adjustment to the data is not necessary. If any $y_i = 0$ we recommend a small adjustment such as

$$\hat{\eta}_i = \log\left(\frac{y_i + 1/2}{t_i} \right) - c_1, \tag{4.2}$$

$$[I(\hat{\eta}_i)]^{-1} = \frac{c_2}{y_i + 1/2} \tag{4.3}$$

where $c_1 > 0$ and $c_2 \geq 1$. The constants c_1 and c_2 represent location and scale parameters, respectively. Since 1/2 is added to y_i, in the case that any $y_i = 0$, the constant c_1 has the

effect of shifting the distribution to its original location. The constant c_2 can be chosen to increase the variance.

One goal of this work is to provide a method to perform Bayesian data analysis of GLMs that requires modest programming effort by the data analyst. The conditionals needed to implement the Gibbs sampler are standard distributions and can be sampled from directly. We ran a Markov chain with 1000 iterations. Index plots of samples drawn from the posterior distributions of the hyperparameters appear in Figures 2, 3, and 4. These plots represent correlated sample drawn from the posterior distributions. Point estimates are calculated from the means of these samples. Burn-in is not a significant issue in this example. The fitted regression line along with the data are shown in Figure 5. One observation appears as an outlier in the design space, but it is not influential since very little exposure is associated with it.

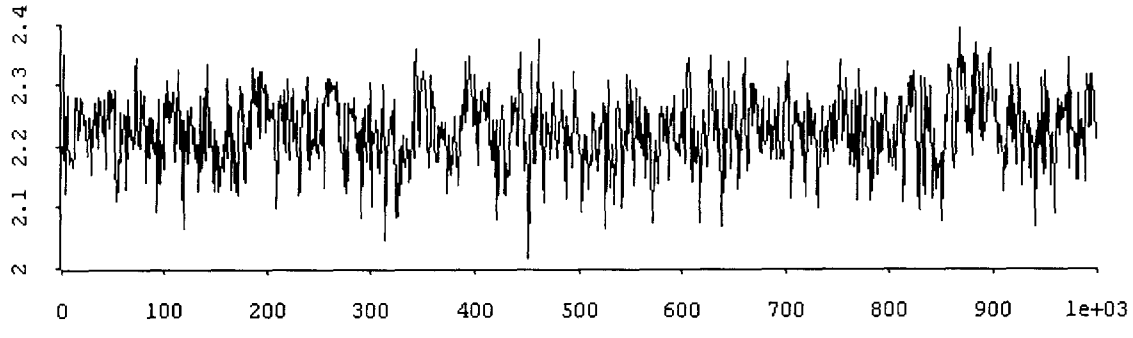

Figure 2. Index Plot of 1000 observations drawn from $\beta_0|y$ with $E[\beta_0 \mid y] = 2.227$

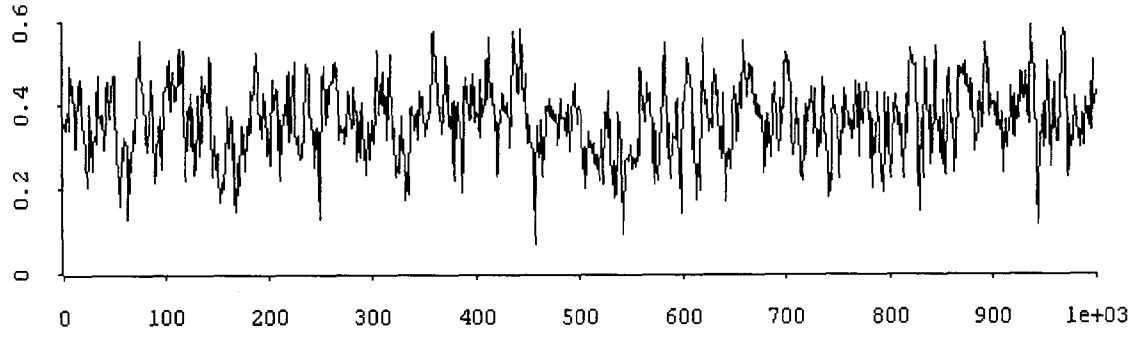

Figure 3. Index Plot of 1000 observations drawn from $\beta_1|y$ with $E[\beta_1 \mid y] = 0.368$

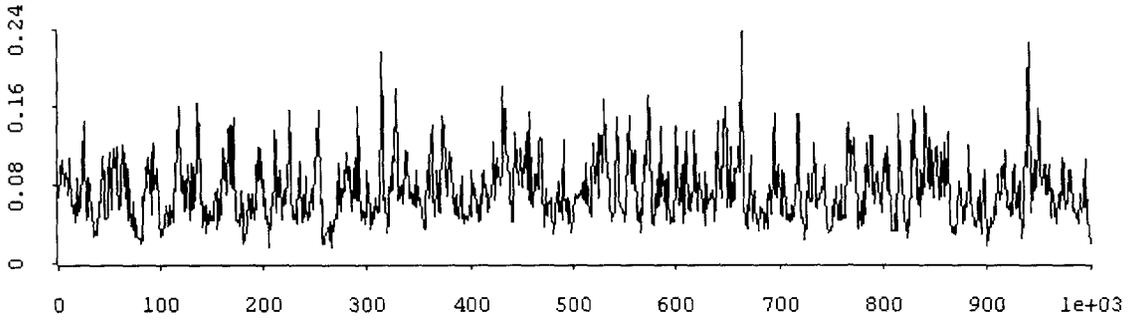

Figure 4. Index Plot of 1000 observations drawn from $\sigma^2|y$ with $E[\sigma^2 \mid y] = 0.077$

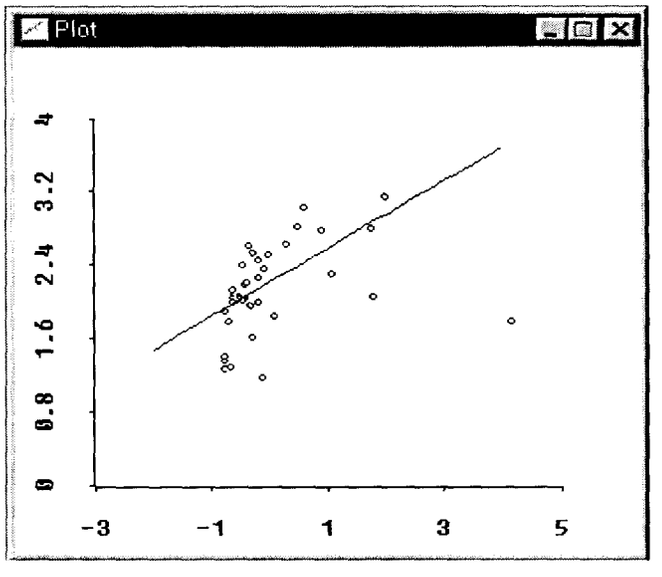

Figure 5. Scatter plot of data and fitted regression line

5. Discussion

A general method for Bayesian data analysis of GLMs has been presented that uses a normal approximation to the likelihood on the scale of the link function. A normal distribution then becomes the conjugate prior and all normal model theory becomes available. We choose the usual noninformative hyperprior for brevity, but there is no need to do so. The extension to a proper prior is straightforward.

Note that in Equation (3.2) it is possible to marginalize over η to obtain the posterior

distribution of the hyperparameters. We choose not to do so to avoid the unequal variance case since the marginal distribution of y on the transformed scale, conditional on the hyperparameters, becomes normal with nonconstant variance. In addition, by sampling from the conditional of η, a sample is generated from the linear predictor and a simple transformation provides case-level samples from GLM parameters.

Finally, this work addresses Bayesian methods for the class of GLMs, but the methods presented here can be applied very generally outside this class. Future work is planned in that direction.

References

[1] Agresti, A. (1990). Categorical Data Analysis. Wiley.

[2] Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85 398-409.

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

[4] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice.* Chapman and Hall.

[5] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, 2nd. edn. Chapman and Hall.

[6] Naylor, J. C. and Smith, A.F.M. (1982). Applications of a method for the efficient computation of posterior distributions. *Appl. Statist.* 31 214-225.

[7] Rosenkranz, S.L. and Raftery, A.E. (1994). Covariate selection in hierarchical models of hospital admission counts: a Bayes factor approach. Technical report no. 268, Department of Statistics, University of Washington.

[8] Schmeiser, B. and Chen, M.H. (1991). General hit-and-run Monte Carlo sampling for evaluating multidimensional integrals. Technical report, School of Industrial Engineering, Purdue University.

[9] Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1996). *BUGS: Bayesian inference Using Gibbs Sampling, Versian 0.5.* Cambridge: Medical Research Council Biostatistics Unit.

[10] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist* 22 1701-1762.

[11] Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginals. *J. Amer. Statist. Assoc.* 81 82-86.

[12] Tierney, L., Kass, R.E., and Kadane, J.B. (1989). Approximate marginal densities for nonlinear functions. *Biometrika* 76 425-434.