The Korean Communications
in Statistics Vol. 11 No. 2, 2004
pp. 381-397

Marginal Likelihoods for Bayesian Poisson Regression Models[1]

Hyunjoong Kim[2], Balgobin Nandram[3], Seong-Jun Kim[4], Ilsu Choi[5], Yunkee Ahn[6], and ChulEung Kim[7]

Abstract

The marginal likelihood has become an important tool for model selection in Bayesian analysis because it can be used to rank the models. We discuss the marginal likelihood for Poisson regression models that are potentially useful in small area estimation. Computation in these models is intensive and it requires an implementation of Markov chain Monte Carlo (MCMC) methods. Using importance sampling and multivariate density estimation, we demonstrate a computation of the marginal likelihood through an output analysis from an MCMC sampler.

Keywords : Poisson regression, Metropolis-Hastings sampler, multivariate density estimation, importance sampler

1. Introduction

The marginal likelihood is now an important tool in Bayesian model selection and model averaging. The computation of the marginal likelihood has attracted considerable interest in the recent Markov chain Monte Carlo (MCMC) literature. A recent and very comprehensive review is given by Han and Carlin (2001). In this article, we address the problem of computing the marginal likelihood for selecting a model.

Let M_1 and M_2 be two models, and let $\boldsymbol{d}$ be a vector of observations. The model specifies

1) This work was supported by Yonsei University Research Fund of 2003
2) Assistant Professor, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea
E-mail: hkim@yonsei.ac.kr
3) Professor, Department of Mathematical Sciences, Worcester Polytechnic Institute, MA 01605, U.S.A.
4) Associate Professor, Department of Industrial Engineering, Kangnung National University, Kangnung 210-702, Korea
5) Associate Professor, Department of Applied Mathematics, Yosu National University, Yeosu 550-749, Korea
6) Professor, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea
7) Associate Professor, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea

a structure for $\boldsymbol{d}|\Omega$ with a proper prior on $P(\Omega)$. Then the marginal likelihood for M_k, denoted by $M_k(\boldsymbol{d})$, is given by $M_k(\boldsymbol{d}) = P(\boldsymbol{d}|M_k) = \int P(\boldsymbol{d}|\Omega, M_k)P(\Omega), k = 1, 2.$

If $P(\boldsymbol{d}|M_1)$ is larger than $P(\boldsymbol{d}|M_2)$, M_1 is preferred to M_2. The usefulness of the marginal likelihood is associated with the Bayes factor which is used in Bayesian hypothesis testing problems. In fact, the Bayes factor is $P(\boldsymbol{d}|M_2)/P(\boldsymbol{d}|M_1)$. This is a measure of strength of the evidence provided by the data for M_2 relative to M_1; see Kass and Raftery (1995). As an application, one can use the marginal likelihoods to select the best model within a set of candidate models. The model with the largest marginal likelihood is the best.

Much work has also been done on the direct estimation of the marginal likelihood in general non-nested model setting (Chib (1995); Gelfand and Dey (1994)) and on the estimation of ratios of marginal likelihoods especially in the setting of nested models (Chen and Shao (1998); DiCiccio, Kass, Raftery and Wasserman (1997); Meng and Wong (1996); Verdinelli and Wasserman (1995)).

2. Computing Marginal Likelihood

Chib (1995) suggested an approach to compute marginal likelihood from the Gibbs sampler output. But these methods work when the posterior conditional densities have simple forms. For generalized linear models Chib and Jeliazkov (2001) extended the method of Chib (1995) to obtain an approach to compute marginal likelihood from the Metropolis-Hastings (M-H) sampler output. Nandram and Kim (2002) simplified the method by using the multiplication rule of probability to exploit the hierarchical structure of models.

In this article, we address the problem of computing marginal likelihood for Poisson regression models. Although the method is applicable generally, we choose to discuss it using small area estimation where we first encountered this problem (see Nandram (2000) for a review). We describe the negative marginal quasi log-likelihoods for two specific models that are currently used for mortality data analysis and disease mapping.

Chib (1995) suggested an approach to compute marginal likelihood from the Gibbs sampler output. It is well known that once the posterior distribution $P(\Omega|\boldsymbol{d})$ is available, by Bayes' theorem, we have

$$M(\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\Omega)P(\Omega)}{P(\Omega|\boldsymbol{d})}. \tag{2.1}$$

Chib (1995) noticed that $M(\boldsymbol{d})$ in (2.1) is invariant to choices of Ω; thus we can use any Ω value for our convenience, but he correctly suggested a high density point. One natural choice is the posterior mode and a simpler choice is the posterior mean which can be easily obtained from an output analysis of any Markov chain Monte Carlo sampler. For generalized linear model, (2.1) is difficult to compute because of non-conjugacy. He has shown how to do this for the probit model, and because he used latent variables, the problem of non-conjugacy disappears.

It is important for our work that even though $M(\boldsymbol{d})$ in (2.1) is not defined if $P(\Omega)$ is improper, we can still find the value of $M(\boldsymbol{d})$ provided $P(\Omega|\,\boldsymbol{d})$ is proper. When the $P(\Omega)$ is not proper but $P(\Omega|\,\boldsymbol{d})$ is proper, we use $Q(\boldsymbol{d}) = -\log(M(\boldsymbol{d}))$, called negative marginal quasi log-likelihood (NMQL), to rank the models.

We describe how to obtain the marginal likelihood and the negative marginal quasi log-likelihood for Poisson regression models.

3. Poisson Regression Models

3.1 A Class of Generalized Linear Model

Let d_{ij} denote a non-negative discrete random variable, and n_{ij} be the sample size, fixed by a design, $d_{ij} \le n_{ij}$, $i=1,\cdots,N, j=1,\cdots,c.$. We assume that d_{ij} given θ_{ij} are independent with

$$f(d_{ij}|\theta_{ij}) = \exp\{P(\theta_{ij}; d_{ij}, n_{ij})\} \tag{3.1}$$

where θ_{ij} are unknown parameters. We assume that there are covariates, $x_{kij}, k=1,\cdots,p-1$ and $\boldsymbol{x} = (1, x_{1ij},\cdots,x_{(p-1)ij})^T$. We also assume that there is a one-to-one function $g(\cdot)$ such that

$$g(\theta_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \nu_i + \delta_j \tag{3.2}$$

where ν_i and δ_i are random effects. For the random effects we take

$$\nu_i|\gamma^2 \sim iid\ N(0,\gamma^2) \text{ and } \delta_j|\sigma^2 \sim iid\ N(0,\sigma^2). \tag{3.3}$$

The specification (3.3) induces a "borrowing of strength" which is desirable, and indeed a popular idea in the small area estimation.

For the hyper-parameters $\boldsymbol{\beta}, \gamma^2, \sigma^2$, we take

$$\boldsymbol{\beta} \sim N(\beta_o, \Delta_o) \text{ or } P(\boldsymbol{\beta}) = 1 \tag{3.4}$$

and

$$\gamma^{-2}, \sigma^{-2} \sim iid\ \Gamma(\eta_o/2, \xi_o/2) \tag{3.5}$$

where $\beta_o, \Delta_o, \eta_o$ and ξ_o are to be specified. For our purpose, we can tolerate a flat prior in (3.4) so long as the posterior distribution is proper. Also with no prior information in (3.5), it is standard practice to take $\eta_o = \xi_o = 0.002$.

The model specifications in (3.1)-(3.5) form a class of generalized linear models. If $d_{ij} | \theta_{ij} \sim ind$ Poisson $(n_{ij}\theta_{ij})$,

$$P(\theta_{ij}, d_{ij}, n_{ij}) = d_{ij}\log(\theta_{ij}) - n_{ij}\theta_{ij} + d_{ij}\log(n_{ij}) - \log(d_{ij}!).$$

Then, $g(\theta_{ij})$ is taken to be the natural parameter of this one-parameter exponential family, $g(\theta_{ij}) = \log(\theta_{ij})$ and the model in (3.1)-(3.5) is called a hierarchical Bayesian Poisson regression model.

3.2 Descriptions of Two Models

We consider two Poisson regression models in the discussion which are popular in small area estimation problem. Both models have the standard specification

$$d_{ij} | \lambda_{ij} \sim ind \text{ Poisson}(n_{ij}\lambda_{ij}),\ i = 1, \cdots, N,\ j = 1, \cdots, c. \tag{3.6}$$

Typically for mortality data, d_{ij} is the number of deaths, n_{ij} is the population sizes, λ_{ij} is the age specific mortality rate in health service area i and age class j.

As the first model (Model 1), we take the link function to be

$$\log\lambda_{ij} = \boldsymbol{x}_j^T\underline{\beta} + v_i \text{ and } v_i | \sigma^2 \sim iid\ N(0, \sigma^2)$$

where $i = 1, \cdots, N$ and $j = 1, \cdots, c$. The covariate $\boldsymbol{x}_j$ is set to denote the age classes. We take a locally uniform prior distribution on $\underline{\beta}$ and a proper diffuse prior on σ^2,

$$P(\underline{\beta}) = 1 \text{ and } \sigma^{-2} \sim \Gamma(\eta_o/2, \xi_o/2) \text{ where } \eta_o = \xi_o = 0.002.$$

This model, called the offset model, is commonly used for data analysis in small area estimation. The joint posterior density for this model is

$$\begin{aligned} P(\underline{\beta}, \underline{v}, \sigma^2 \,|\, \boldsymbol{d}) \propto\ & \exp\Big[\sum_{i=1}^{N}\sum_{j=1}^{c}\{(\boldsymbol{x}_j^T\underline{\beta} + v_i)d_{ij} - n_{ij}\exp(\boldsymbol{x}_j^T\underline{\beta} + v_i)\}\Big] \\ \times\ & \sigma^{-N}\exp\Big\{\sum_{i}^{N} - v_i^2/2\sigma^2\Big\} \times (\sigma^{-2})^{\eta_o/2-1}\exp\{-\xi_o/2\sigma^2\}. \end{aligned} \tag{3.7}$$

As a second model (Model 2), we take the link function to be

$$\log(\lambda_{ij}) = \boldsymbol{x}_j^T\underline{\beta} + \nu_i + \delta_j$$

where

$$\nu_i|\gamma^2 \sim iid\ N(0,\gamma^2) \text{ and } \delta_j|\sigma^2 \sim iid\ N(0,\sigma^2)$$

and

$$P(\underline{\beta}) = 1 \text{ and } \gamma^{-2},\sigma^{-2} \sim iid\ \Gamma(\eta_o/2,\xi_o/2), \eta_o = \xi_o = 0.002.$$

It is anticipated that δ_{ij} can accommodate extra variation. The joint posterior density of this model is

$$\begin{aligned} P(\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2\Big| \boldsymbol{d}) \propto\ & \exp\Big[\sum_{i=1}^{N}\sum_{j=1}^{c}\{(\boldsymbol{x}_j^T\underline{\beta}+\nu_i+\delta_j)d_{ij} - n_{ij}\exp(\boldsymbol{x}_j^T\underline{\beta}+\nu_i+\delta_j)\}\Big] \\ \times\ & \gamma^{-N}\exp\Big\{\sum_{i=1}^{N}(-\nu_i^2/2\gamma^2)\Big\}\times\sigma^{-c}\exp\Big\{\sum_{j=1}^{c}(-\delta_j^2/2\sigma^2)\Big\} \\ \times\ & (\gamma^{-2})^{\eta_o/2-1}\exp\{-\xi_o/2\gamma^2\}\times(\sigma^{-2})^{\eta_o/2-1}\exp\{-\xi_o/2\sigma^2\}. \end{aligned} \tag{3.8}$$

It is convenient to make the transformation

$$\phi_j = \boldsymbol{x}_j^T\underline{\beta} + \delta_j$$

keeping all others untransformed. Then the joint posterior density is

$$\begin{aligned} P(\underline{\beta},\underline{\nu},\underline{\phi},\gamma^2,\sigma^2\Big| \boldsymbol{d}) \propto\ & \exp\Big[\sum_{i=1}^{N}\sum_{j=1}^{c}\{(\nu_i+\phi_j)d_{ij} - n_{ij}\exp(\nu_i+\phi_j)\}\Big] \\ \times\ & \gamma^{-N}\exp\Big\{\sum_{i=1}^{N}(-\nu_i^2/2\gamma^2)\Big\}\times\sigma^{-c}\exp\Big[\sum_{j=1}^{c}\{-(\phi_j-\boldsymbol{x}_j^T\underline{\beta})^2/2\sigma^2\}\Big] \\ \times\ & (\gamma^{-2})^{\eta_o/2-1}\exp\{-\xi_o/2\gamma^2\}\times(\sigma^{-2})^{\eta_o/2-1}\exp\{-\xi_o/2\sigma^2\}. \end{aligned} \tag{3.9}$$

The gain by the transformation is that the conditional posterior density of $\underline{\beta}$ is multivariate normal; see Gelfand, Sahu and Carlin (1995) for more details. We obtained samples from the joint posterior density in (3.9) by using the Metropolis-Hastings sampler.

3.3 Negative Marginal Quasi Log-likelihood

For Model 1,

$$M_1(\boldsymbol{d}) = \frac{P(\boldsymbol{d}|\underline{\beta},\underline{\nu},\sigma^2)P(\underline{\beta},\underline{\nu},\sigma^2)}{P(\underline{\beta},\underline{\nu},\sigma^2|\boldsymbol{d})}$$

where

$$P(\boldsymbol{d}\mid\underline{\beta},\underline{\nu},\sigma^2)=\prod_{i=1}^{N}\prod_{j=1}^{c}\left[n_{ij}^{d_{ij}}\exp\{(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)d_{ij}-n_{ij}\exp(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)\}\Big/d_{ij}!\right] \tag{3.10}$$

$$P(\underline{\beta},\underline{\nu},\sigma^2)=\left[\prod_{i=1}^{N}(2\pi\sigma^2)^{-1/2}\exp\{-\nu_i^2/2\sigma^2\}\right] \times(\xi_o/2)^{\eta_o/2}(\sigma^{-2})^{\eta_o/2-1}\exp\{-\xi_o/2\sigma^2\}/\Gamma(\eta_o/2) \tag{3.11}$$

and $P(\underline{\beta},\underline{\nu},\sigma^2\mid\boldsymbol{d})$ in (3.7) is the joint posterior density function. Here we let $Q_1(\boldsymbol{d})=-\log(M_1(\boldsymbol{d}))$, and we evaluate (3.10) and (3.11) and the posterior density at the posterior means of $\Omega\mid\boldsymbol{d}$ where $\Omega=(\underline{\beta},\underline{\nu},\sigma^2)$.

Now

$$P(\underline{\beta},\underline{\nu},\sigma^2\mid\boldsymbol{d})=P(\underline{\nu}\mid\underline{\beta},\sigma^2,\boldsymbol{d})P(\underline{\beta},\sigma^2\mid\boldsymbol{d}). \tag{3.12}$$

Observe that

$$P(\underline{\nu}\mid\underline{\beta},\sigma^2,\boldsymbol{d})=\frac{\prod_i\left\{\prod_j\exp\{(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)d_{ij}-n_{ij}\exp(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)\}\right\}\exp\{-\nu_i^2/2\sigma^2\}}{\int_{R^N}\left[\prod_i\left\{\prod_j\exp\{(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)d_{ij}-n_{ij}\exp(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)\}\right\}\exp\{-\nu_i^2/2\sigma^2\}d\underline{\nu}\right.}$$

where $\underline{\nu}=(\nu_1,\cdots,\nu_N)^T\in R^N$.

We show how to evaluate the normalization constant

$$I=\int_{R^N}\left[\prod_i\left\{\prod_j\exp\{(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)d_{ij}-n_{ij}\exp(\boldsymbol{x}_j^T\underline{\beta}+\nu_i)\}\right\}\exp\{-\nu_i^2/2\sigma^2\}\right]d\underline{\nu}$$

in Appendix B.

We use density estimation with a multivariate normal kernel to evaluate $P(\underline{\beta},\sigma^2\mid\boldsymbol{d})$; see Appendix A. To obtain a more symmetric density, we transform $\underline{\beta},\sigma^2$ to $\underline{\beta},\tau$ where $\tau=\log\sigma^2$. With this transformation

$$P(\underline{\beta},\sigma^2\mid\boldsymbol{d})=P(\underline{\beta},\tau\mid\boldsymbol{d})_{\{\tau=\log(\sigma^2)\}}/\sigma^2.$$

Thus, we apply multivariate density estimation to get $P(\underline{\beta},\tau\mid\boldsymbol{d})$ evaluated at the posterior mean of τ.

For Model 2,

$$M_2(\boldsymbol{d})=\frac{P(\boldsymbol{d}\mid\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2)P(\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2)}{P(\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2\mid\boldsymbol{d})}$$

where

$$P(\boldsymbol{d}\mid\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2\sigma^2)=\prod_{i=1}^{N}\prod_{j=1}^{c}\left[n_{ij}^{d_{ij}}\exp\{(\boldsymbol{x}_j^T\underline{\beta}+\nu_i+\delta_j)d_{ij}-n_{ij}\exp(\boldsymbol{x}_j^T\underline{\beta}+\nu_i+\delta_j)\}\Big/d_{ij}!\right], \tag{3.13}$$

$$P(\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2)=\left[\prod_{i=1}^{N}(2\pi\gamma^2)^{-1/2}\exp\{-\nu_i^2/2\gamma^2\}\right]\left[\prod_{i=1}^{N}(2\pi\sigma^2)^{-1/2}\exp\{-\delta_i^2/2\sigma^2\}\right], \quad (3.14)$$
$$\times(\xi_o/2)^{\eta_o}(\gamma\sigma)^{-\eta_o+2}\exp\{-\xi_o(1/\gamma^2+1/\sigma^2)/2\}/\ \Gamma(\eta_o/2)^2$$

and $P(\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2|\,\boldsymbol{d})$ in (3.8) is the joint posterior density function. Here we let $Q_2(\boldsymbol{d})=-\log(M_2(\boldsymbol{d}))$, and we evaluate (3.13) and (3.14) and the posterior density at the posterior means of $\Omega|\,\boldsymbol{d}$ where $\Omega=(\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2)$.

Consider the transformation

$$\phi_j=\boldsymbol{x}_j^T\underline{\beta}+\delta_j,\quad j=1,\cdots,c$$

with identity on the other parameters. We have

$$P(\underline{\beta},\underline{\nu},\underline{\delta},\gamma^2,\sigma^2|\,\boldsymbol{d})=P(\underline{\beta},\underline{\nu},\underline{\phi},\gamma^2,\sigma^2|\,\boldsymbol{d})_{\{\phi_j=\ \boldsymbol{x}_j^T\underline{\beta}+\delta_j,\,j=1,\cdots,c\}}$$

and we need to evaluate $P(\underline{\beta},\underline{\nu},\underline{\phi},\gamma^2,\sigma^2|\,\boldsymbol{d})$. Now,

$$P(\underline{\beta},\underline{\nu},\underline{\phi},\gamma^2,\sigma^2|\,\boldsymbol{d})=P(\underline{\beta}|\underline{\nu},\underline{\phi},\gamma^2,\sigma^2,\boldsymbol{d})P(\underline{\nu}|\underline{\phi},\gamma^2,\sigma^2,\boldsymbol{d})P(\underline{\phi},\gamma^2,\sigma^2|\,\boldsymbol{d}) \quad (3.15)$$

where

$$\underline{\beta}|\underline{\nu},\underline{\phi},\gamma^2,\sigma^2,\boldsymbol{d}\ \sim\ N\left\{(\sum_j\boldsymbol{x}_j\boldsymbol{x}_j^T)^{-1}\sum_j\phi_j\boldsymbol{x}_j,\sigma^2(\sum_j\boldsymbol{x}_j\boldsymbol{x}_j^T)^{-1}\right\}$$

and

$$P(\nu|\underline{\phi},\gamma^2,\sigma^2,\boldsymbol{d})=\frac{\prod_i\{\prod_j\exp\{(\nu_i+\phi_j)d_{ij}-n_{ij}\exp(\nu_i+\phi_j)\}\}\exp\{-\nu_i^2/2\gamma^2\}}{\int_{R^N}\prod_i\{\prod_j\exp\{(\nu_i+\phi_j)d_{ij}-n_{ij}\exp(\nu_i+\phi_j)\}\}\exp\{-\nu_i^2/2\gamma^2\}d\underline{\nu}}.$$

We evaluate

$$\int_{R^N}\left[\prod_i\{\prod_j\exp\{(\nu_i+\phi_j)d_{ij}-n_{ij}\exp(\nu_i+\phi_j)\}\}\exp\{-\nu_i^2/2\gamma^2\}\right]d\underline{\nu}$$

in a manner similar to (3.12) for Model 1; see Appendix B where we simply set $\phi_j\equiv\boldsymbol{x}_j^T\underline{\beta}$. We summarize this method in Appendix C.

For $P(\underline{\phi},\gamma^2,\sigma^2|\,\boldsymbol{d})$ we use the multivariate density estimation. We transform $\underline{\phi}=\underline{\phi}$, $\tau_1=\log\gamma^2$, $\tau_2=\log\sigma^2$ to get

$$P(\underline{\Phi}, \gamma^2, \sigma^2 | \boldsymbol{d}) = P(\underline{\Phi}, \tau_1, \tau_2 | \boldsymbol{d}) \{_{\tau_1 = \log(\gamma^2), \tau_2 = \log(\sigma^2)}\} / \gamma^2 \sigma^2$$

where again the transformation is used for symmetrization.

4. Some Numerical Examples

In Section 4.1, we review two alternative measures which we compare with the NMQL. We employ these three measures to discriminate between the two models described in Section 3. We use an example on colon cancer in Section 4.2 and a small scale simulation study in Section 4.3, to compare the three measures and the two models.

4.1 Two Alternative Measures

The first alternative method of evaluating the models is to use a cross-validation. Let $\boldsymbol{d}_{(ij)}$ denote the set of all data $\boldsymbol{d}$'s except for (ij). Then letting $r_{ij} = d_{ij} / n_{ij}$, we define the cross-validation residual as $a_{ij} = r_{ij} - E(r_{ij} | \boldsymbol{d}_{(ij)})$, and the standardized cross-validation residual as

$$DRES_{ij} = a_{ij} / SD(r_{ij} | d_{ij}). \tag{4.1}$$

That is, the (ij)-th observed r_{ij} is "held out" and compared with its point estimator, $E(r_{ij} | \boldsymbol{d}_{(ij)})$, which is evaluated without using the observed d_{ij}. We use (4.1), in summary form, to rank the two models, and we employ the cross-validation residuals as a measure of concordance of the data with a proposed model. For simplicity we count the number of health service areas with $| DRES_{ij} | \geq 3$ for all i and j, and we call this quantity NHD3.

The second alternative method of evaluating the models is to use the posterior expected predictive deviance (EPD),

$$E\{P(\boldsymbol{d}^{obs}, \boldsymbol{d}^{new}) | \boldsymbol{d}^{obs}\} \tag{4.2}$$

where $\boldsymbol{d}^{new}$ is a random vector with distribution

$$f(\boldsymbol{d}^{new} | \boldsymbol{d}^{obs}) = \int g(\boldsymbol{d}^{new} | \underline{\lambda}) h(\underline{\lambda} | \boldsymbol{d}^{obs}) d\underline{\lambda} \tag{4.3}$$

with $h(\underline{\lambda} | \boldsymbol{d}^{obs})$ the posterior density of $\underline{\lambda}$ and $g(\boldsymbol{d}^{new} | \underline{\lambda})$ the probability mass function of

$\boldsymbol{d}^{new}$ in (3.6). In (4.2), $P(\boldsymbol{d}^{obs}, \boldsymbol{d}^{new})$ is a measure of the discrepancy between $\boldsymbol{d}^{obs}$, the observed vector of the d_{ij}, and $\boldsymbol{d}^{new}$, a set of "new" observations. We select $\boldsymbol{d}^{new}$ from the posterior predictive distribution of $\boldsymbol{d}^{new}$ in (4.3). If the model and data are concordant, $\boldsymbol{d}^{obs}$ and $\boldsymbol{d}^{new}$ should be similar and (4.2) should be small. We use the Poisson-based measure $P(\cdot, \cdot)$,

$$P(\boldsymbol{d}^{obs}, \boldsymbol{d}^{new}) = 2\sum_{i}^{N}\sum_{j}^{c}\{(d_{ij}^{obs}+0.5)\log\{(d_{ij}^{obs}+0.5)/(d_{ij}^{new}+0.5)\} - (d_{ij}^{obs}-d_{ij}^{new})\}.$$

See for example Waller, Carlin, Xia and Gelfand(1997) and Gelfand and Ghosh(1998).

4.2 Example on Colon Cancer

Colon cancer is one of the diseases of the middle age and the elderly. We use mortality data for white males with colon cancer collected 1988-1992 for 6 regions of the U.S. In column 2 of <Table I> we present the number of health service areas in each of the 6 regions. We apply Models 1 and 2 to these data. There are 7 age classes in the data. The covariate $\boldsymbol{x}$ in the models is used to describe the age effect.

We fitted both models using the Metropolis-Hastings algorithm. In each case we "burn in" 1000 iterates and picked every 20th thereafter to get 1000 iterates which we use for model assessment and inference.

In <Table I>, we compare the two models using the three measures. NHD3 indicates that Model 2 fits better than Model 1 for all regions except region 7 (NHD3=2 for Model 1 versus NHD3=5 for Model 2). The EPD shows that Model 2 performs better than Model 1 in all regions. According to NMQL Model 2 is better in all regions.

<Table I> NHD3, expected predictive deviance, and negative marginal quasi loglikelihood (NMQL) for Models 1 and 2 by region.

		NMQL		NHD3		Poisson-based EPD	
Region	#of HSA's	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
1	23	995	821	7	2	567	310
2	121	4456	4054	32	11	2415	1632
3	45	1711	1468	5	4	638	517
4	105	3672	3135	13	6	1519	1222
5	38	1567	1197	2	5	513	424
6	48	1642	1183	20	3	1149	661

4.3 A Small-scale Simulation Study

Our example on colon cancer in Section 4.2 indicates that the three measures we used to rank Models 1 and 2 are consistent. We investigate this consistency further using a small-scale simulation study.

We generated a data set similar to the observed data on colon cancer. Using Model 2, we estimated β, γ^2, and σ^2 by an output analysis from the Metropolis-Hastings sampler on the observed data. We kept β fixed at its posterior mean, and obtained the median values of γ^2, and σ^2, denoted by $\widehat{\gamma^2}$, and $\widehat{\sigma^2}$, respectively. We used a 3^2 design with γ^2 at three levels $\left(\frac{1}{2}\widehat{\gamma^2}, \widehat{\gamma^2}, 2\widehat{\gamma^2}\right)$ and σ^2 at three levels $\left(\frac{1}{2}\widehat{\sigma^2}, \widehat{\sigma^2}, 2\widehat{\sigma^2}\right)$, and we generated the data for each region by taking

$$\nu_i|\gamma^2 \sim iid\ N(0,\gamma^2),\ \delta_j|\sigma^2 \sim iid\ N(0,\sigma^2) \text{ and } d_{ij}|\lambda_{ij} \sim ind \text{ Poisson}(n_{ij}\lambda_{ij})$$

where $\log(\lambda_{ij}) = x_j^T\beta + \nu_i + \delta_j$ as for colon cancer. That is, we generated nine dataset from Model 2, and we fitted both Model 1 and Model 2 to each of the nine simulated data sets as described for the data on colon cancer.

In <Table Ⅱ>, we present the three measures for the low value of γ^2. For the low value of σ^2, the three measures are consistent for region 4 where Model 2 is worse than Model 1 by far especially for NMQL. The three measures show consistently that Model 2 is better than Model 1 for all other regions. For the median and high values of σ^2, the degree of consistency among the three measures remains the same except for a reversal in the NMQL for regions 4 in favor of Model 2.

In <Table Ⅲ>, the three measures for the median value of γ^2 are presented. Again, Model 2 is preferred by all three measures. For the low value of σ^2, the three measures are consistent. For the median and high values of σ^2, NMQL prefers model 2 at region 4 while NHD3 and EPD fail to favor model 2.

In <Table Ⅳ>, we present the three measures for the high value of γ^2. For the low value of σ^2, the three measures are consistent. For the median values of σ^2, region 4 is inconsistent but NMQL moves to the correct direction that prefers model 2. Model 2 is favored consistently for the high value of σ^2.

In general, the three measures show a high degree of consistency. As σ^2 increases, we expect that Model 2 will show better performance because Model 1 does not have the random effects δ_j. For NMQL, we observed this clearly. For region 4, as the value of σ^2 increases from low to medium (or high), NMQL selects model 2. However, both NHD3 and EPD fail to choose model 2 over model 1. We observe that NMQL is a sensible quantity to use for ranking models.

<Table II> Negative marginal quasi log-likelihood (NMQL), NHD3, and expected predictive deviance for Models 1 and 2 using simulated data by three levels of δ^2 when γ^2 is low.

γ^2 =Low		NMQL		NHD3		Poisson-based EPD	
δ^2	Region	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Low	1	2569	2388	10	1	654	342
	2	9861	9482	58	1	4367	1728
	3	2994	2683	5	2	745	606
	4	7646	38141	3	9	1400	1518
	5	17748	4908	4	0	622	459
	6	3359	1036	22	0	1652	649
Median	1	4037	3856	15	1	872	342
	2	15977	14605	92	5	6970	1751
	3	4403	3880	10	2	847	584
	4	10012	8571	4	6	1460	1516
	5	28794	9229	13	2	868	507
	6	4903	1047	31	1	2590	679
High	1	7396	6977	16	2	1552	326
	2	29252	25510	108	8	12583	1723
	3	8675	7489	15	4	971	577
	4	16486	14386	6	6	1464	1482
	5	58923	21336	24	1	1383	447
	6	9322	1216	37	2	4742	704

5. Conclusion

We have discussed the Poisson regression models for small area estimation. Then we have shown how to compute the negative marginal quasi log-likelihood when there are improper priors but proper posterior densities. Our method uses importance sampling and a multivariate density estimation from an output analysis of the Metropolis-Hastings sampler.

Using an example on colon cancer and a small-scale simulation study, we have shown that the NMQL agrees reasonably well with two other measures proposed in the literature. This adds credence to the NMQL even though it is not really a marginal likelihood since the prior distributions are improper. However, our methodology applies equally well to the marginal likelihood that is obtained from a proper prior.

<Table III> Negative marginal quasi log-likelihood (NMQL), NHD3, and expected predictive deviance for Models 1 and 2 using simulated data by three levels of δ^2 when γ^2 is the median.

γ^2 =Median		NMQL		NHD3		Poisson-based EPD	
δ^2	Region	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Low	1	2917	2536	9	0	682	335
	2	10538	10148	64	4	4553	1717
	3	2828	2447	7	3	762	605
	4	7762	10561	6	7	1380	1411
	5	17793	16885	4	2	616	497
	6	3708	1116	22	0	1684	684
Median	1	4408	4176	14	0	879	322
	2	16950	15275	90	3	7041	1765
	3	4466	3953	11	6	805	576
	4	11162	9228	6	8	1456	1503
	5	27103	8495	10	4	833	488
	6	5225	1126	31	2	2598	666
High	1	7849	7410	18	1	1697	328
	2	29408	25535	110	5	12485	1742
	3	8642	7707	15	4	987	526
	4	18340	15522	5	7	1465	1491
	5	98014	45600	30	3	1953	498
	6	9543	1330	37	1	4708	618

Appendix A. Multivariate Density Estimation

Let $x_1, x_2, \ldots, x_n$ be a random sample from an unknown k-variate distribution. Let

$$\bar{x} = \sum_{i=1}^{n} x_i/n \quad \text{and} \quad S^2 = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T/(n-1).$$

Then an estimator for the probability density function is $\hat{f}(x)$ where

$$\hat{f}(x) = \frac{1}{n h_{opt}^k [\det(S)]^{1/2}} \sum_{i=1}^{n} K(x - x_i)^T S^{-1}(x - x_i)/h_{opt}^2, \tag{A.1}$$

for any $x \in R^k$. In (A.1), $K(t) = (2\pi)^{-k/2} \exp\{-t/2\}, 0 < t < \infty$, is the kernel and

<Table IV> Negative marginal quasi log-likelihood (NMQL), NHD3, and expected predictive deviance for Models 1 and 2 using simulated data by three levels of δ^2 when γ^2 is high.

γ^2 =High		NMQL		NHD3		Poisson-based EPD	
δ^2	Region	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Low	1	3438	2915	8	0	649	329
	2	11284	10486	56	2	4356	1698
	3	2963	2568	12	1	777	598
	4	9492	45487	2	3	1405	1582
	5	19174	4910	6	3	689	515
	6	4498	1251	20	0	1655	678
Median	1	5047	4392	14	0	1037	345
	2	17572	15685	88	2	6884	1678
	3	4319	4009	10	2	796	587
	4	11950	10278	3	6	1398	1430
	5	28219	8737	7	1	745	464
	6	5477	1287	31	1	2481	687
High	1	8637	7753	18	3	1556	342
	2	30164	27202	110	3	12485	1704
	3	8424	7319	14	6	886	577
	4	19128	16969	9	5	1460	1459
	5	68086	24568	25	6	1627	546
	6	10695	1813	39	2	4939	710

$$h_{opt} = \{4/(k+2)\}^{1/(k+4)} n^{-1/(k+4)}$$

is the optimal window width for the multivariate normal population (Silverman 1986).

In practice, (A.1) works best if the components of x are symmetric or (more optimistically) approximately normally distributed (Silverman 1986}. Thus we apply (A.1) after a degree of symmetrization.

Appendix B. Evaluation of the Normalization Constant in Model 1

We need

$$I = \int_{R^N} [\Pi_{i=1}^N \exp\{\Delta(\nu_i)\}] d\underline{\nu},$$

where

$$\Delta(\nu_i) = \Delta_1(\nu_i) + \Delta_2(\nu_i)$$

with

$$\Delta_1(\nu_i) = \sum_j \left\{ (x_j^T \underline{\beta} + \nu_i) d_{ij} - n_{ij} e\,(x_j^T \underline{\beta} + \nu_i) \right\} \text{ and } \Delta_2(\nu_i) = -\nu_i^2/2\sigma^2. \qquad \text{(B.1)}$$

We use importance sampling in which

$$I = \int_{R^N} \frac{[\Pi_{i=1}^N e\{\Delta(\nu_i)\}]}{p_\eta(\underline{\nu}|\mathrm{d})} p_\eta(\underline{\nu}|\boldsymbol{d}) d\underline{\nu}$$

where $p_\eta(\underline{\nu}|\mathrm{d})$ is an N-variate Student's t distribution on η degrees of freedom with location and scale parameters to be determined.

Observe that we can perform N univariate integrations using Student's t densities but it is more efficient to use just a single integration in R^N. This is true because $\nu_1, \ldots, \nu_N$ given $\underline{\beta}, \sigma^2, \mathrm{d}$ are independent.

We obtain $p_\eta(\underline{\nu}|\boldsymbol{d})$ in the following manner. We note first that

$$\frac{\partial \Delta_1}{\partial \nu_i} = \sum_j \left\{ d_{ij} - n_{ij} \exp(x_j^T \underline{\beta} + \nu_i) \right\} \text{ and } \frac{\partial^2 \Delta_1}{\partial \nu_i^2} = -\sum_j n_{ij} \exp(x_j^T \underline{\beta} + \nu_i).$$

Then setting $\frac{\partial \Delta_1}{\partial \nu_i} = 0$, we have $\hat{\nu}_i^* = \log \left\{ \frac{\sum_j d_{ij}}{\sum_j n_{ij} \exp(x_j^T \underline{\beta})} \right\}$. and approximately

$$\nu_i | \underline{\beta}, \sigma^2, \boldsymbol{d} \sim N(\hat{\nu}_i^*, (\sum_j d_{ij})^{-1}). \qquad \text{(B.2)}$$

Now combining (B.1) and (B.2) we have approximately

$$\nu_i | \underline{\beta}, \sigma^2, \boldsymbol{d} \sim N \left(\frac{(\sum_j d_{ij}) \hat{\nu}_i^*}{1/\sigma^2 + \sum_j d_{ij}}, \frac{1}{1/\sigma^2 + \sum_j d_{ij}} \right).$$

We obtain $p_\eta(\underline{\nu}|\underline{\beta}, \sigma^2, \mathrm{d})$ by using the latent variable ρ^2 with

$$\nu_1, \ldots, \nu_N | \rho^2, \underline{\beta}, \sigma^2, \mathbf{d} \sim ind\ N\left(\frac{(\sum_j d_{ij})\hat{\nu}_i^*}{1/\sigma^2 + \sum_j d_{ij}}, \frac{\rho^2}{1/\sigma^2 + \sum_j d_{ij}} \right) \tag{B.3}$$

and

$$\frac{\eta}{\rho^2} \sim \chi^2_\eta, \tag{B.4}$$

where η is to be specified.

Note that in the univariate case we must draw N values of ρ^2 in (B.4); also $\nu_1, \ldots, \nu_N | \rho^2, \underline{\beta}, \sigma^2, \mathbf{d}$ are now correlated. We draw MN-variate vectors $\underline{\nu}$ in (B.3) and (B.4), denoted by $\underline{\nu}^{(h)}$, $j = 1, \ldots, M$. Then we estimate I by

$$\hat{I} = M^{-1} \sum_{h=1}^{M} \frac{\Pi_{i=1}^N \exp\{\Delta(\nu_i^{(h)})\}}{p_\eta(\underline{\nu}^{(h)} | \underline{\beta}, \sigma^2, \mathrm{d})}. \tag{B.5}$$

We found that $M = 1000$ with $\eta = 10$ is conservative in (B.5).

For Model 2, we repeat (B.1)-(B.5) with $\phi_j \equiv \boldsymbol{x}_j^T \underline{\beta}$.

Appendix C. Evaluation of the Normalization Constant in Model 2

As in Section 3.3, we start with

$$P(\underline{\beta}, \underline{\nu}, \underline{\phi}, \gamma^2, \sigma^2 | \mathrm{d}) = \mathrm{P}(\underline{\nu}, \underline{\phi} | \underline{\beta}, \gamma^2, \sigma^2, \mathrm{d}) \mathrm{P}(\underline{\beta}, \gamma^2, \sigma^2 | \mathrm{d})$$

where $P(\underline{\beta}, \gamma^2, \sigma^2 | \mathrm{d})$ is described in Section 3.3 and $P(\underline{\nu}, \underline{\phi} | \underline{\beta}, \gamma^2, \sigma^2, \mathrm{d})$ is obtained by computing its normalization constant

$$I_a = E\Big[\exp\Big\{ \sum_{i=1}^{N} \sum_{j=1}^{c} \{ d_{ij}(\boldsymbol{x}_j^T \underline{\beta} + \nu_i + \delta_j) - n_{ij} \exp(\boldsymbol{x}_j^T \underline{\beta} + \nu_i + \delta_j) \} - \sum_{i=1}^{N} \nu_i^2 / 2\gamma^2 - \sum_{j=1}^{c} \delta_j^2 / 2\sigma^2 \Big\} \Big/ f_a(\underline{\nu}, \underline{\delta} | \underline{\beta}, \gamma^2, \sigma^2, \boldsymbol{d}) \Big] \tag{C.1}$$

where the expectation is taken over $f_a(\underline{\nu}, \underline{\delta} | \underline{\beta}, \gamma^2, \sigma^2 \mathrm{d})$.

Let $\hat{\underline{\nu}}$ and $\hat{\underline{\delta}}$ be estimates from an output analysis of the Metropolis-Hastings sampler, and let $\hat{\nu}_i^* = \log\left(\sum_{j=1}^{c} d_{ij} \Big/ \sum_{j=1}^{c} n_{ij} \exp\{\boldsymbol{x}_j^T \underline{\beta} + \hat{\delta}_j\} \right)$ and $\hat{\delta}_j^* = \log\left(\sum_{i=1}^{N} d_{ij} \Big/ \sum_{i=1}^{N} n_{ij} \exp\{\boldsymbol{x}_j^T \underline{\beta} + \hat{\nu}_i\} \right)$ where

$i=1,...,N$ and $j=1,...,c$. Then, $f_a(\underline{\nu},\underline{\delta}|\underline{\beta},\gamma^2,\sigma^2,\mathbf{d})$ in (C.1) is obtained by the following construction,

$$f_a(\underline{\nu},\underline{\delta}|\underline{\beta},\gamma^2,\sigma^2,\mathbf{d})=f_a(\underline{\nu}|\hat{\underline{\delta}},\underline{\beta},\gamma^2,\sigma^2,\mathbf{d})f_a(\underline{\delta}|\hat{\underline{\nu}},\underline{\beta},\gamma^2,\sigma^2,\mathbf{d})$$

and after introducing the latent variable τ^2

$$\nu_i|\hat{\underline{\delta}},\underline{\beta},\gamma^2,\sigma^2,d,\tau^2 \sim ind\ N\left(\frac{\hat{\nu}_i^*\sum_j d_j}{1/\gamma^2+\sum_j d_{ij}},\frac{\tau^2}{1/\gamma^2+\sum_j d_{ij}}\right), \tag{C.2}$$

$$\delta_j|\hat{\underline{\nu}},\underline{\beta},\gamma^2,\sigma^2,d,\tau^2 \sim ind\ N\left(\frac{\hat{\delta}_j^*\sum_i d_{ij}}{1/\sigma^2+\sum_i d_{ij}},\frac{\tau^2}{1/\sigma^2+\sum_i d_{ij}}\right) \tag{C.3}$$

and

$$\eta/\tau^2 \sim \chi^2_\eta \tag{C.4}$$

where $\eta=10$, $i=1,...,N$, and $j=1,...,c$. The construction in (C.2)-(C.4) produces a cN-variate Student's t density for $f_a(\underline{\nu},\underline{\delta}|\underline{\beta},\gamma^2,\sigma^2,\mathbf{d})$.

References

[1] Chen, M.-H. and Shao, Q.-M. (1998). On Monte Carlo methods for estimating ratios of normalizing constants, *Annals of Statistics* **57**, 1563-1594.

[2] Chib, S. (1995). Marginal likelihood from the Gibbs output, *Journal of the American Statistical Association* **90**, 1313-1321.

[3] Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output, *Journal of the American Statistical Association* **96**, 270-281.

[4] DiCiccio, T. J., Kass, R. E., Raftery, A. E. and Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations, *Journal of the American Statistical Association* **92**, 903-915.

[5] Gelfand, A. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations, *Journal of the Royal Statistical Society*, **B56**, 501-514.

[6] Gelfand, A. and Ghosh, S. (1998). Model choice: Minimum posterior predictive loss approach, *Biometrika* **85**, 1-11.

[7] Gelfand, A., Sahu, S. and Carlin, B. (1995). Efficient parameterizations for normal linear mixed models, *Biometrika* **82**, 479-488.

[8] Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors: A comparative review, *Journal of the American Statistical Association* **96**, 1122-1132.

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

[10] Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration, *Statistica Sinica* **6**, 831-860.

[11] Nandram, B. (2000). Bayesian generalized linear models for inference about small areas, in D. K. Dey, S. K. Ghosh and B. K. Mallick (eds), *Generalized linear models: A Bayesian Perspective*, Marcel Dekker, pp. 91-114.

[12] Nandram, B. and Kim, H. (2002). Marginal likelihood for a class of bayesian generalized linear models, *Journal of Statistical Computation and Simulation* **72**, 319-340.

[13] Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[14] Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalizationof the Savage-Dickey density ratio, *Journal of the American Statistical Association* **90**, 614-618.

[15] Waller, L., Carlin, B., Xia, H. and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates, *Journal of the American Statistical Association* **92**, 607-617.

[Received November 2003, Accepted July 2004]