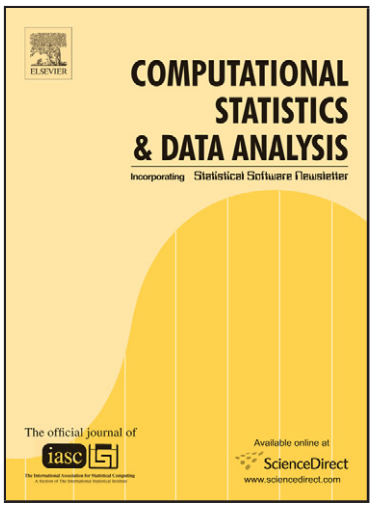
Efficient and accurate approximate Bayesian inference with an application to insurance data

Author's Accepted Manuscript



www.elsevier.com/locate/csda

Efficient and accurate approximate Bayesian inference with an application to insurance data

George Streftaris, Bruce J. Worton

PII: S0167-9473(07)00342-8
DOI: doi:10.1016/j.csda.2007.09.006
Reference: COMSTA 3807

To appear in: *Computational Statistics & Data Analysis*

Received date: 5 December 2006
Revised date: 7 September 2007
Accepted date: 8 September 2007

Cite this article as: George Streftaris and Bruce J. Worton, Efficient and accurate approximate Bayesian inference with an application to insurance data, *Computational Statistics & Data Analysis* (2007), doi:10.1016/j.csda.2007.09.006

Efficient and accurate approximate Bayesian inference with an application to insurance data

George Streftaris [a,*], Bruce J. Worton [b]

[a] *School of Mathematical and Computer Sciences and Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, U.K.*

[b] *School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh, James Clerk Maxwell Building, King's Buildings, Mayfield Road, Edinburgh EH9 3JZ, U.K.*

Abstract

Efficient and accurate Bayesian Markov chain Monte Carlo methodology is proposed for the estimation of event rates under an overdispersed Poisson distribution. An approximate Gibbs sampling method and an exact independence-type Metropolis–Hastings algorithm are derived, based on a log-normal/gamma mixture density that closely approximates the conditional distribution of the Poisson parameters. This involves a moment matching process, with the exact conditional moments obtained employing an entropy distance minimisation (Kullback-Liebler divergence) criterion. A simulation study is conducted and demonstrates good Bayes risk properties and robust performance for the proposed estimators, as compared with other estimating approaches under various loss functions. Actuarial data on insurance claims are used to illustrate the methodology. The approximate analysis displays superior Markov chain Monte Carlo mixing efficiency, whilst providing almost identical inferences to those obtained with exact methods.

Key words: Bayes risk; Entropy distance; Effective sample size; Hierarchical Bayesian analysis; Insurance claims; Markov chain Monte Carlo; Mixture distribution; Monte Carlo error

* Corresponding author. Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Colin Maclaurin Building, Heriot-Watt University, Edinburgh EH14 4AS, U.K. Tel: +44 131 451 3679; Fax: +44 131 451 3249
Email address: G.Streftaris@hw.ac.uk (George Streftaris).

1 Introduction

Simultaneous inference for several Poisson distributions has attracted much attention, especially in the case of additional variation caused by the dependence among the Poisson parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\mathrm{T}}$. Applications involving such inference have emerged in various areas, including actuarial science (Haberman and Renshaw, 1996; Makov et al., 1996) and epidemiology (Clayton and Kaldor, 1987; Ainsworth and Dean, 2006). The problem has been tackled in the past using various shrinkage estimating approaches, aiming to exploit the information provided in the entire vector of the Poisson parameters $\boldsymbol{\theta}$ (e.g. Morris, 1983). Bayesian methodology provides a natural framework for exploiting the relation between the components of $\boldsymbol{\theta}$ through the prior distribution, thus also dealing with the problem of overdispersion. The use of Bayes and empirical Bayes methods for the analysis of Poisson data in actuarial science, including the consideration of Poisson/gamma models for insurance claims, is discussed by Makov et al. (1996), Haastrup (2000), Czado et al. (2005) and Ntzoufras et al. (2005) among others. Advances in Markov chain Monte Carlo (MCMC) stochastic integration methodology (e.g. Tierney, 1994) have facilitated the generic implementation of full Bayesian analysis in related problems.

In this paper we work under a hierarchical Bayesian framework assuming a log-normal prior distribution for the Poisson parameters, and develop very efficient and accurate MCMC methodology for posterior analysis. The motivation for the work is that generally, with non-conjugate models, efficiency in the mixing behaviour of a Markov chain usually depends on the choice and construction of suitable proposal distributions, and this is often not given sufficient consideration in easily implemented MCMC algorithms employed in the analysis of hierarchical models. We propose a method which improves on the efficiency of commonly used Gibbs and Metropolis–Hastings schemes, while retaining the accuracy of the posterior inference. The presented approach may also be extended to a larger class of related models, where other prior distributions are assumed for the Poisson parameters.

We investigate the use of a close approximation to the conditional distribution of the Poisson rates $\theta_1, \theta_2, \ldots, \theta_m$, given all other model parameters and the data. The proposed approximation is based on a log-normal/gamma mixture density which matches the first three moments of the original distribution. For the computation of the moments of the posterior distribution we use a method relying on entropy distance (Kullback-Liebler divergence) minimisation. The resulting density is then employed in a Gibbs sampling scheme which mixes more efficiently than traditionally used approaches, and provides very accurate posterior inference that also performs favourably in terms of Bayes risk. We also employ the approximate density as the pro-

posal distribution for an independence-type Metropolis–Hastings step (Tierney, 1994), which gives a very efficient exact algorithm with acceptance rate close to 1. In contrast, standard MCMC approaches often incorporate Gibbs sampling algorithms using rejection sampling techniques (e.g. as implemented in the *WinBUGS* software, http://www.mrc-bsu.cam.ac.uk/bugs), or various Metropolis–Hastings schemes that rely on fine-tuning the variance of candidate distributions (Gelfand and Smith, 1990; George et al., 1994; Damien et al., 1999). Although the implementation of such approaches appears to be straightforward, mixing efficiency is not necessarily guaranteed. Our results demonstrate that ease of implementation can be offset by an increase in the number of iterations required to achieve a certain level of estimation precision, which can be important in problems with slow chain mixing. Despite the fact that rapid advance of computing power continuously changes the balance among the developing, computing and running time of algorithms, there is arguably still scope for improved efficiency.

The proposed methodology is applied to the estimation of insurance claim intensities. The analysis confirms the accuracy of the methodology, and additionally demonstrates its convergence efficiency in terms of mixing of the Markov chain.

In Section 2 we introduce the Poisson/log-normal hierarchical model, while the derivation of the MCMC methods for the analysis is outlined in Sections 3–5. The results of an extensive Monte Carlo simulation study are presented in Section 6 to assess the risk properties of the proposed estimators, and compare them with those of other Bayes and classical methods under various scenarios concerning the prior distribution and loss function. The insurance data application is discussed in Section 7, where we also compare the efficiency of the considered MCMC schemes.

2 The model

We assume that given the parameters $\theta_1, \theta_2, \ldots, \theta_m$, the counts $Y_1, Y_2, \ldots, Y_m$, are conditionally independent Poisson random variables with respective means $\theta_i E_i,\ i = 1, \ldots, m$, i.e.

$$Y_i|\theta_i \sim \text{Poisson}(\theta_i E_i), \qquad i = 1, \ldots, m, \tag{1}$$

where $E_i,\ i = 1, \ldots, m$, represent different exposure times. The parameters $\theta_i,\ i = 1, \ldots, m$, give the rate of occurrence of events and depend on p explanatory variables in a log-linear regression structure expressed as

$$\log(\theta_i) = \mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \varepsilon_i, \qquad i = 1, \ldots, m, \tag{2}$$

where $\mathbf{x}_i^{\mathrm{T}} = (x_{1,i}, x_{2,i}, \ldots, x_{p,i})$, for $i = 1, \ldots, m$, are known values of the explanatory variables, $\mathbf{b} = (b_1, b_2, \ldots, b_p)^{\mathrm{T}}$ is a vector of regression coefficients and ε_i, $i = 1, \ldots, m$, are random error terms.

In actuarial science (1) and (2) can be used to model the occurrence of insurance claims. In this context, Y_i represents the number of actual claims in group $i = 1, \ldots, m$, $\mathbf{x}_i$ are covariates related to group i (e.g. age), and E_i is the total time of exposure of group i to a specific policy. Overdispersion in observed insurance claims data occurs due to a number of duplicate policies among policy holders (Currie and Waters, 1991). For example, a classical generalised linear model (GLM) analysis on the number of claims data described in Section 7 (and presented in Table 4), reveals some extra-Poisson variation (GLM deviance 17.2 on 6 degrees of freedom). The use of GLMs in actuarial science is discussed in detail by Haberman and Renshaw (1996). In the simple exchangeable model, the maximum likelihood estimator for θ_i is given as $\hat{\theta}_i^{\mathrm{ML}} = \frac{Y_i}{E_i}$, $i = 1, \ldots, m$. Although $\hat{\theta}_i^{\mathrm{ML}}$ is the minimum variance unbiased estimator, it is inadmissible under various loss functions when two or more conditionally independent Poisson distributions are involved (Hudson, 1978). This is because it ignores the remaining components of the data vector, which are important in situations where the estimation of each individual element of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\mathrm{T}}$ benefits from the information incorporated in the entire parameter vector.

Under a full hierarchical Bayes framework, we assume a log-normal prior distribution for the event rates $\theta_i, i = 1, \ldots, m$ to allow for overdispersion in the data. In addition to (1), the error terms ε_i, $i = 1, \ldots, m$, in (2) are assumed to be identically and independently distributed as $N(0, \sigma^2)$ random variables, or equivalently the parameters θ_i, $i = 1, \ldots, m$, are independently distributed according to a log-normal distribution $\mathrm{LN}\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{b}, \sigma^2\right)$, with mean $\exp\left(\mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \frac{1}{2}\sigma^2\right)$ and variance $\exp(2\mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \sigma^2)(e^{\sigma^2} - 1)$. If we let λ_i denote the natural logarithm of θ_i, then we can write

$$\lambda_i = \log(\theta_i) \sim N(\mathbf{x}_i^{\mathrm{T}}\mathbf{b}, \sigma^2), \qquad i = 1, \ldots, m. \tag{3}$$

At the second prior stage we assume that the hyperprior parameters $\mathbf{b}$ and σ^2 are jointly distributed according to the flat uniform prior density

$$\pi(\mathbf{b}, \sigma^2) \propto 1, \tag{4}$$

reflecting vague prior information. This improper prior can be defined as the limit of independent proper distributions: $b_k \sim N(0, \sigma_{b_k}^2)$, as $\sigma_{b_k}^2 \to \infty$; and $\tau = 1/\sigma^2 \sim \mathrm{Pareto}(1, r_{\sigma^2}^{-1})$, as $r_{\sigma^2} \to \infty$. The latter is equivalent to $\sigma^2 \sim U(0, r_{\sigma^2})$, and is commonly considered in Bayesian analysis (e.g. Gelman and Rubin, 1992; Spiegelhalter et al., 1996; O'Hagan and Forster, 2004). Alternatively, the inverse-gamma prior $\sigma^2 \sim \text{Inv-Ga}(\alpha_{\sigma^2}, \beta_{\sigma^2})$ is often assumed,

with $\alpha_{\sigma^2} \to 0$ and $\beta_{\sigma^2} \to 0$ providing a non-informative improper prior. At the limit we obtain $\pi(\sigma^2) \propto \sigma^{-2}$, that is a uniform prior for the logarithm of σ^2. This prior often leads to a non-integrable posterior density in normal models (Berger, 1985; Gelman et al., 2004). In our model, as applied to the insurance data in Section 7, the posterior distribution of σ^2 is sensitive to the choice of very small values for the hyperparameters α_{σ^2} and β_{σ^2}, reflecting the infinite peak of $\pi(\sigma^2) \propto \sigma^{-2}$ in the region close to 0, and thus suggesting that this prior may not always be regarded as non-informative in the class of models examined here. We also note that the increasing weight attached to the prior as $\sigma^2 \to 0$, can lead to over-smoothed Poisson estimates, as a result of excessive shrinking towards the prior mean. This is often not desirable in problems where infrequent small or zero counts should not be overlooked (as for example in the analysis of insurance claims or spatial modelling of disease occurrence). O'Hagan and Forster (2004, p.311) discuss the necessity for careful consideration of the prior on variance parameters when different weak priors lead to differences in posterior inferences.

The prior distribution in (3) can also take other forms, defining in the general case a model which is also known in actuarial science as a compound Poisson sampling model (Carlin and Louis, 2000), or a mixed Poisson model (e.g. Grandell, 1997).

3 Bayesian inference

MCMC estimation requires the conditional distribution for each model parameter, given all other parameters and the data. In terms of the parameterisation $\lambda_i = \log(\theta_i)$, $i = 1, \ldots, m$, the joint posterior density of the parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathrm{T}}$, $\mathbf{b}$ and σ^2 is given by

$$p(\boldsymbol{\lambda}, \mathbf{b}, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}m} \exp\left[\sum_{i=1}^{m} \left\{ \lambda_i y_i - E_i e^{\lambda_i} - \frac{1}{2}\sigma^{-2}(\lambda_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{b})^{\mathbf{2}} \right\}\right]. \quad (5)$$

Consider the full conditional distribution for the regression coefficients b_k, $k = 1, \ldots, p$. Define $\mathbf{b}_{-k}$ as the vector $\mathbf{b}$ with its kth component omitted, that is $\mathbf{b}_{-k} = (b_1, \ldots, b_{k-1}, b_{k+1}, \ldots, b_p)^{\mathrm{T}}$, and denote the corresponding ith linear component by $h_{-k,i}$, i.e.

$$h_{-k,i} = b_1 x_{1,i} + \ldots + b_{k-1} x_{k-1,i} + b_{k+1} x_{k+1,i} + \ldots + b_p x_{p,i}. \quad (6)$$

Then, using the decomposition $\mathbf{x}_i^{\mathrm{T}}\mathbf{b} = h_{-k,i} + b_k x_{k,i}$ the full conditional distri-

bution of b_k given $\boldsymbol{\lambda}$, $\mathbf{b}_{-k}$, σ^2 and $\mathbf{y}$ is given from (5) as

$$b_k|\boldsymbol{\lambda}, \mathbf{b}_{-k}, \sigma^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^m(\lambda_i - h_{-k,i})x_{k,i}}{\sum_{i=1}^m x_{k,i}^2},\ \frac{\sigma^2}{\sum_{i=1}^m x_{k,i}^2}\right), \tag{7}$$

for $k = 1, \ldots, p$. The joint posterior density in (5) also implies that the conditional posterior distribution of σ^2 given $\boldsymbol{\lambda}, \mathbf{b}$ is

$$\sigma^2 \ \Big|\ \boldsymbol{\lambda}, \mathbf{b}, \mathbf{y} \sim \text{Inv-Ga}\left(\frac{m-2}{2}, \frac{\sum_{i=1}^m(\lambda_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{b})^2}{2}\right). \tag{8}$$

Clearly, this distribution is valid when $m > 2$, which is the case in practice in all problems related to the one discussed here. Using $\pi(\sigma^2) \propto \sigma^{-2}$ changes the shape parameter of the above distribution to $\frac{m}{2}$, which does not alter the algorithms proposed in the remaining of the paper.

As far as the full conditional posterior distribution of the parameter vector $\boldsymbol{\lambda}$ is concerned, (5) shows that given $\mathbf{b}$ and σ^2, the parameters $\lambda_1, \ldots, \lambda_m$, have independent full conditional densities given by

$$p(\lambda_i|\mathbf{b}, \sigma^2, \mathbf{y}) \propto \exp\left\{\lambda_i y_i - E_i e^{\lambda_i} - \frac{1}{2}\,\sigma^{-2}(\lambda_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{b})^2\right\}, \qquad i = 1 \ldots, m. \tag{9}$$

Therefore, simulating from the full conditionals of $b_1, \ldots, b_p$, and σ^2 is straightforward using normal and inverse gamma distributions. However, as is evident from the form of (9), this is not the case for the full conditional distribution of λ_i, and thus we derive a method for simulating from this distribution in the following sections.

4 Approximation to the conditional distribution of the event rates

The full conditional density of $\theta_i = e^{\lambda_i}$ may be expressed as

$$p(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y}) \propto \theta_i^{y_i - 1}\ e^{-E_i\theta_i}\ \exp\left\{-\frac{\left(\log\theta_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{b}\right)^2}{2\sigma^2}\right\}, \tag{10}$$

for $i = 1, \ldots, m$. The density in (10) involves a gamma and a log-normal density and is considerably skewed. To approximate it we propose a flexible mixture of these two components, which is very accurate for appropriately selected parameter values. Experimentation with either the gamma or the log-normal density alone, demonstrated that these approximations are not able to sufficiently capture the skewness of $p(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})$. Therefore, the proposed

approximating density is of the form

$$f = \rho f_{LN} + (1-\rho) f_{Ga}, \tag{11}$$

where $0 \leq \rho \leq 1$, f_{LN} and f_{Ga} denote a mixing parameter and the log-normal density and gamma density components respectively. The density f is calculated by matching its mean, variance and skewness to those of the full conditional distribution. Specifically, the parameters of a $\mathrm{LN}(\delta, \tau^2)$ and a $\mathrm{Ga}(a, b)$ distribution are chosen so that the mean and the variance of each of these two distributions are equal to the mean and the variance of the original full conditional distribution. This gives

$$\delta = \log\{\mathrm{E}(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})\} - \frac{\tau^2}{2}, \qquad \tau^2 = \log\left\{1 + \frac{\mathrm{var}(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})}{\mathrm{E}^2(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})}\right\}, \tag{12}$$

for the log-normal part of the mixture distribution, and

$$a = \frac{\mathrm{E}^2(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})}{\mathrm{var}(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})}, \qquad b = \frac{\mathrm{E}(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})}{\mathrm{var}(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})}, \tag{13}$$

for the parameters of the gamma distribution. Using the parameters in (12) and (13), the third order moments about zero for the log-normal and gamma distributions can be computed as $\mu'_{3,LN} = \exp(3\delta + \frac{9}{2}\tau^2)$ and $\mu'_{3,Ga} = a(a+1)(a+2)/b^3$ respectively. Then, with the first two moments matched, the mixing proportion ρ is determined in a way such that the mixture distribution also has the same skewness as the full conditional. This is achieved by equating the third order moments of the two distributions (Titterington et al., 1985, p.72), i.e. $\mu'_3 = \rho\mu'_{3,LN} + (1-\rho)\mu'_{3,Ga}$, where μ'_3 denotes the third moment about zero of the full conditional distribution. Simulating from the log-normal/gamma mixture is straightforward and simply involves sampling from a $\mathrm{LN}(\delta, \tau^2)$ distribution with probability ρ and from a $\mathrm{Ga}(a, b)$ with probability $1-\rho$.

For the mixture proposal (11) to be a good approximation to the full conditional density of θ_i, the method requires accurate computation of the moments of the full conditional distribution involved in (12) and (13). Below we outline an efficient method for achieving this.

4.1 Entropy distance minimisation

The moments of the full conditional distribution of θ_i may be conveniently determined by using entropy distance minimising methodology. If we let $L(\lambda_i|y_i)$ and $\pi(\lambda_i)$ denote the likelihood and prior density for λ_i, and $p(y_i)$ the marginal

density of a single observation y_i, we may write

$$\begin{aligned}\mathrm{E}\left(\theta_i^r|\mathbf{b},\sigma^2,\mathbf{y}\right) &= \int_{-\infty}^{\infty} e^{\lambda_i r}\frac{L(\lambda_i|y_i)\pi(\lambda_i)}{p(y_i)}\,d\lambda_i \\ &= \frac{E_i^{y_i}}{y_i!\;p(y_i)}\int_{-\infty}^{\infty}\exp\left\{\lambda_i(y_i+r)-E_i e^{\lambda_i}\right\}\pi(\lambda_i)\,d\lambda_i \\ &= \frac{(y_i+r)!}{y_i!\;E_i^r}\;\frac{p(y_i+r)}{p(y_i)}. \end{aligned} \tag{14}$$

This demonstrates that we can obtain the approximate conditional posterior moments by deriving an approximation to the marginal density of the data. Such a density can be determined so that it minimises the entropy distance between the joint density of (y_i,λ_i) and an approximation of the form

$$p^*(y_i,\lambda_i) = p^*(y_i)\;p^*(\lambda_i|y_i), \tag{15}$$

where $p^*(y_i)$ approximates the marginal density of y_i and $p^*(\lambda_i|y_i)$ is a normal $N(\alpha_i,\omega_i^2)$ density. The entropy distance (also referred to as the Kullback-Liebler divergence) can be used as a measure of discrepancy between a distribution and an approximation (O'Hagan and Forster, 2004), and expressed in terms of the parameters α_i and ω_i^2 is given by

$$\begin{aligned}\mathcal{D}I(\alpha_i,\omega_i^2) &= \mathrm{E}\left\{\log\frac{p^*(y_i,\lambda_i)}{p(y_i,\lambda_i)}\right\} = \mathrm{E}\left\{\log\frac{p^*(y_i)\;p^*(\lambda_i|y_i)}{L(\lambda_i|y_i)\;\pi(\lambda_i)}\right\} \\ &= \mathrm{E}\left\{\log p^*(y_i) - \log\left(\frac{E_i^{y_i}}{y_i!}\right) + \log(\sigma) - \log(\omega_i)\right. \\ &\qquad \left. -y_i\lambda_i + E_i e^{\lambda_i} - \frac{1}{2}\;\omega_i^{-2}(\lambda_i-\alpha_i)^2 + \frac{1}{2}\;\sigma^{-2}(\lambda_i-\mathbf{x}_i^{\mathrm{T}}\mathbf{b})^2\right\}. \end{aligned}$$

Evaluating expectations with respect to the $N(\alpha_i,\omega_i^2)$ distribution we obtain

$$\begin{aligned}\mathcal{D}I(\alpha_i,\omega_i^2) &= \log p^*(y_i) - \log\left(\frac{E_i^{y_i}}{y_i!}\right) + \log(\sigma) - \log(\omega_i) \\ &\qquad - y_i\alpha_i + E_i e^{\alpha_i+\frac{1}{2}\omega_i^2} - \frac{1}{2} + \frac{1}{2}\;\sigma^{-2}\left\{\omega_i^2 + (\alpha_i-\mathbf{x}_i^{\mathrm{T}}\mathbf{b})^2\right\}, \end{aligned} \tag{16}$$

and the marginal density that sets (16) equal to zero is given by

$$\begin{aligned}p^*(y_i) = \tilde{\omega}_i\;\sigma^{-1}\;E_i^{y_i}(y_i!)^{-1}\exp\left(\frac{1}{2}+\tilde{\alpha}_i y_i - E_i e^{\tilde{\alpha}_i+\frac{1}{2}\tilde{\omega}_i^2}\right) \\ \times\exp\left[-\frac{1}{2}\;\sigma^{-2}\left\{\tilde{\omega}_i^2+(\tilde{\alpha}_i-\mathbf{x}_i^{\mathrm{T}}\mathbf{b})^2\right\}\right], \end{aligned} \tag{17}$$

where $\tilde{\alpha}_i$ and $\tilde{\omega}_i^2$ are the values that minimise (16) for any given $p^*(y_i)$, as shown in Appendix A. Finally, the above density can be used in (14) to provide the required moments.

This approximation may not be sufficiently accurate for our purpose if $y_i = 0$, as the normal approximations involved in the method can be problematic in the presence of zero counts (e.g. Leonard and Hsu, 1999). Therefore, we use an alternative approach for such cases, in which a discretisation of the full conditional density (10) is employed, with the Poisson likelihood multiplied by a discrete approximation to the normal prior distribution of $\lambda_i = \log(\theta_i)$. This prior approximation relies on matching the first 10 moments of the normal prior and discrete distributions, using the approach described in Appendix B. If we let p_j, $j = 1, \ldots, l$, denote the probabilities of a discrete approximation to a standard normal distribution evaluated at the points γ_j, $j = 1, \ldots, l$, and allow θ_i to take the values $\theta_{ij} = \exp(\mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \gamma_j\sigma)$, then θ_{ij} and p_j, $j = 1, \ldots, l$, define an l-point discrete approximation to the prior distribution of θ_i. If we also let q_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, l$, denote the posterior probabilities for θ_{ij}, then Bayes' theorem implies that these may be approximated by

$$q_{ij} = \frac{\theta_{ij}^{y_i} e^{-\theta_{ij}} p_j}{\sum_{j=1}^{l} \theta_{ij}^{y_i} e^{-\theta_{ij}} p_j},$$

for $i = 1, \ldots, m$, $j = 1, \ldots, l$. Thus, an approximation to the rth order conditional moment $\mathrm{E}(\theta_i^r|\mathbf{b}, \sigma^2, \mathbf{y})$ may be computed directly as $\sum_{j=1}^{l} \theta_{ij}^r q_{ij}$, avoiding expensive simulation.

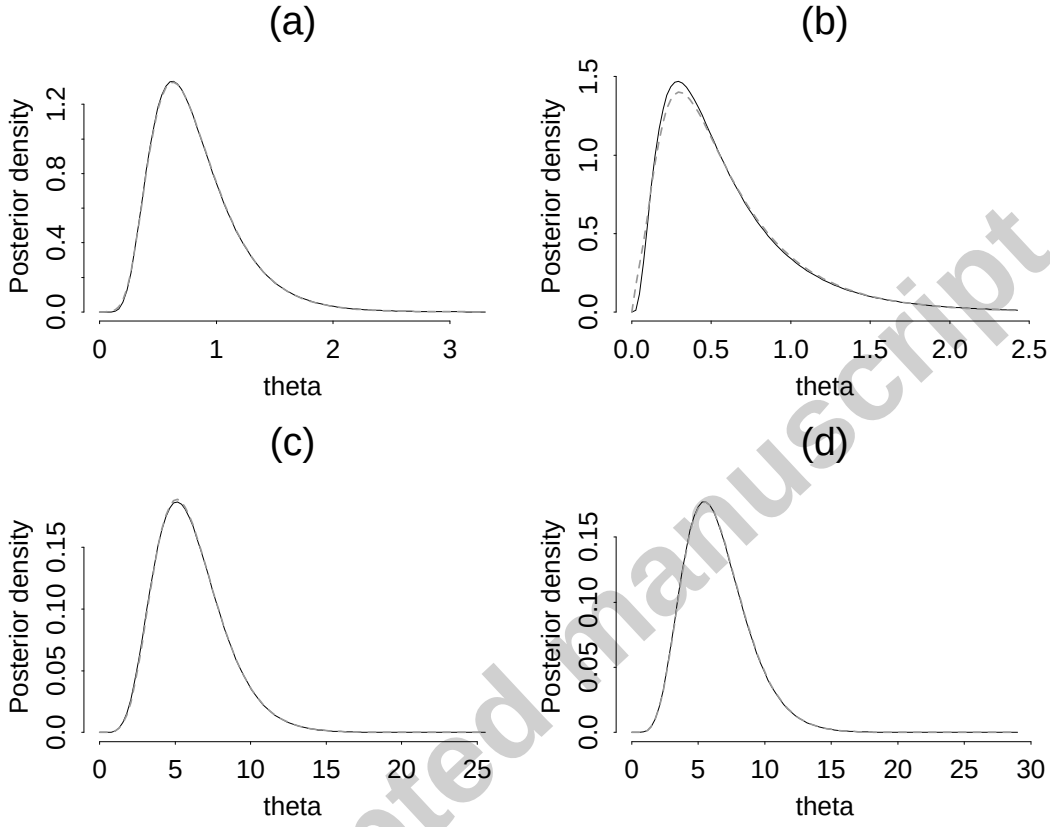
The approximation produced when we employ the strategy outlined in this section is shown in Fig. 1. The mixture approximation is almost indistinguishable from the exact density, which has been computed with expensive numerical integration. Even in the case of Fig. 1b, where $y_i = 0$ and the variation of θ_i is large, the approximation is very good.

5 MCMC schemes based on the proposed approximation

5.1 Approximate Gibbs sampler

The simulation of $b_1, \ldots, b_p$ and σ^2 is easy using standard techniques, as noted in Section 3. Now, from the results of Section 4, we use the mixture density (11) as an approximation to the full conditional distribution of $\theta_i, i = 1, \ldots, m$, allowing the implementation of a Gibbs sampling algorithm which results in accurate posterior analysis. Therefore, in summary, at each iteration of the algorithm we can employ Gibbs steps to sample from the full conditional distributions of $b_1, \ldots, b_p$, σ^2 as well as $\theta_i, i = 1, \ldots, m$, from (7), (8) and the mixture distribution given in (11) respectively.

Fig. 1. Mixture approximation (dashed line) to $p(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})$ (solid line) with various y_i and $(\mathbf{b}, \sigma^2)$ values. The latter determine the coefficient of variation (cv) of the log-normal prior distribution of θ_i: (a) $y_i = 0, \text{cv} = 0.5$; (b) $y_i = 0, \text{cv} = 1.0$; (c) $y_i = 8, \text{cv} = 1.5$; (d) $y_i = 8, \text{cv} = 2.0$. The approximation is produced using the approach of Section 4. The exact full conditional is computed from (10) with numerical integration.



5.2 Independence Metropolis–Hastings chain

In addition to the above approximate scheme, an exact analysis can easily be obtained by using a single-component Metropolis–Hastings algorithm, in which the log-normal/gamma mixture approximation to $p(\theta_i|\mathbf{b}, \sigma^2, \mathbf{y})$ will serve as the proposal density for updating the Poisson parameters in a hybrid MCMC strategy. This forms an independence Metropolis–Hastings chain (Tierney, 1994), since the proposal distribution of θ_i at the current iteration does not depend on the previous iteration value of θ_i, and vice versa. The acceptance probability of the chain involves terms that can be viewed as ratios of the target density and the importance function in an importance sampling scheme, and therefore the acceptance probability is expressed as a ratio of the so-called 'importance weights' (Geweke, 1989). The use of the gamma/log-normal approximation to (10) as the proposal distribution ensures that the acceptance ratio is always approximately 1, as verified with the application in

Section 7 and various other examples using real and simulated data.

The risk performance and efficiency of both approaches described above are compared to other common methods in Sections 6 and 7. It is demonstrated that they are superior to a standard Metropolis-within-Gibbs algorithm that employs a normal proposal density which is matched to the full conditional distribution. However, the approximate Gibbs sampling chain is expected to mix more efficiently than the Metropolis–Hastings algorithm, since it totally avoids rejections. This offers a convergence advantage to the approximate method, as demonstrated by the quantitative comparisons in Section 7.

6 Simulation study of risk properties

We investigate the performance of the approximate hierarchical Bayes (AHB) method and the exact mixture-proposal independence Metropolis–Hastings algorithm described in Section 5, by assessing the performance of the estimators in terms of their Bayes risk

$$\mathrm{E}_\theta \mathrm{E}_{Y|\theta}\{L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})\}, \tag{18}$$

under a loss function $L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta})$, where $\widehat{\boldsymbol{\theta}}$ denotes an estimator of $\boldsymbol{\theta}$. Here we consider the special case in which the exposures $E_i, i = 1, \ldots, m$, in (1) are all equal to 1, and $\lambda_i = \log(\theta_i) \sim N(\eta, \sigma^2)$, where η is a scalar. We consider the squared error loss function

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} (\widehat{\theta_i} - \theta_i)^2, \tag{19}$$

and the normalised squared error loss function

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \frac{(\widehat{\theta_i} - \theta_i)^2}{\theta_i}. \tag{20}$$

Additionally, to investigate the behaviour of the Bayes estimators when the averaging over the m components of the vector parameter $\boldsymbol{\theta}$ is ignored, we employ the maximum component squared error loss function

$$L(\widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \max_{1 \le i \le m} \{(\widehat{\theta_i} - \theta_i)^2\}. \tag{21}$$

For comparison purposes we also evaluate the risk of the hierarchical Bayes (HB) estimator resulting from a Metropolis-within-Gibbs algorithm which uses a normal proposal distribution with mean and variance equal to those of the full conditional, two empirical Bayes (EB) estimators which are based on a linear shrinkage rule, and the maximum likelihood estimator (MLE). The first

EB estimator is derived in a way such that it minimises the Bayes risk among all linear estimators of the same form (Efron and Morris, 1973), and is given by

$$\widehat{\theta}_i^{\,\mathrm{EB}} = (1-c)\; y_i + c\; \bar{y}, \tag{22}$$

where $\bar{y}$ denotes the sample mean of the data, and $c \in [0,1]$ is given as $\frac{\mathrm{E}(\theta_i)}{\mathrm{var}(\theta_i)+\mathrm{E}(\theta_i)}$ and is estimated in the EB context by $\min\{\frac{(m-1)\bar{y}}{\sum_{i=1}^{m}(y_i-\bar{y})^2}, 1\}$. The second EB estimator multiplies the coefficient c by a factor of $\frac{m-3}{m-1}$, and is a modification to (22) proposed by Morris (1983) to resemble the shrinkage behaviour of the HB estimator. Candel (2006) investigates the risk properties of similar EB estimators in a multilevel normal analysis, while Haastrup (2000) compares HB and EB estimators in a Poisson model for actuarial data.

For the simulation study we first generated a number of m true θ_i values, independently from a log-normal distribution having a specified mean $\mathrm{E}(\theta_i)$ and variance $\mathrm{var}(\theta_i)$. Two different values, $m = 10$ and $m = 30$ were used. For each $\theta_i, i = 1, \ldots, m$, a random variate Y_i was then drawn from a Poisson(θ_i) distribution, and the estimates of θ_i, were computed. The Bayes risk (18) was then estimated for each of the loss functions in (19)–(21) by

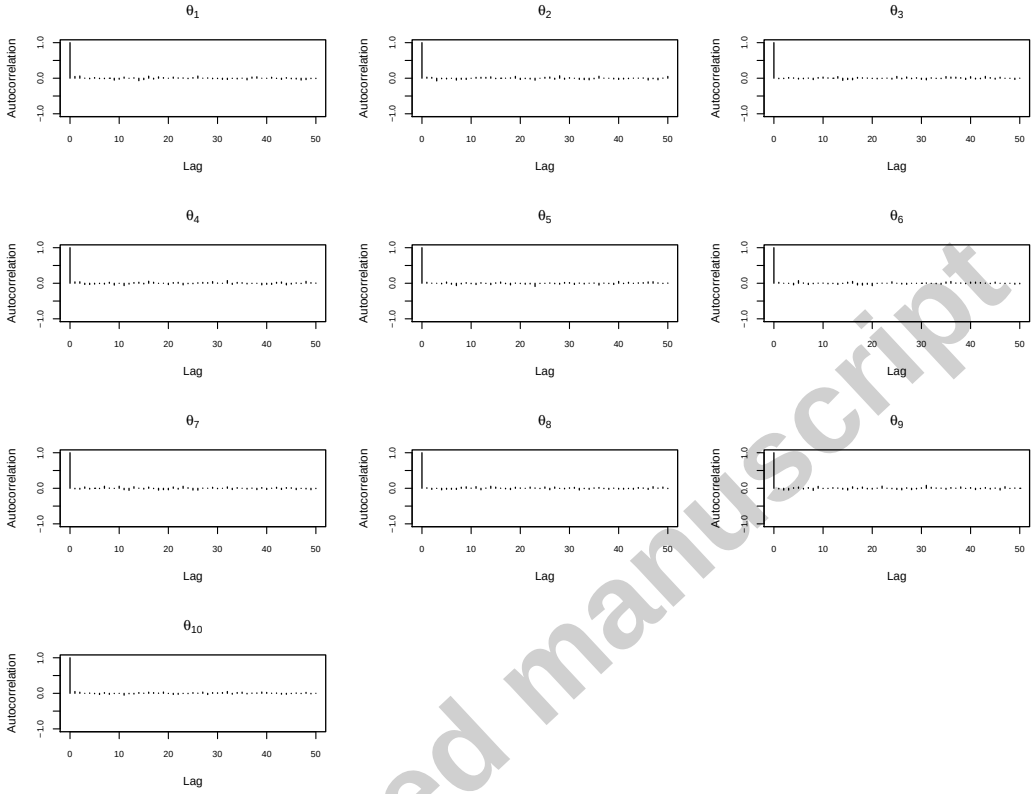
$$\frac{1}{N}\sum_{t=1}^{N}\left\{L(\widehat{\boldsymbol{\theta}}_t, \boldsymbol{\theta}_t)\right\}, \tag{23}$$

where $\widehat{\boldsymbol{\theta}}_t$ is the estimator of $\boldsymbol{\theta}_t$ at repetition t, $t = 1, \ldots, N$. The entire procedure was repeated for various different true $\mathrm{E}(\theta_i)$ and $\mathrm{var}(\theta_i)$ combinations. In the MCMC algorithm for the hierarchical Bayes estimators we employed 1200 simulated values, with the first 200 used as a burn-in. Although this relatively small number of MCMC updates does not necessarily guarantee convergence, inspection of the trace and autocorrelation of the chain did not reveal any problems. Relevant plots using the AHB method are presented in Fig. 2 and show good mixing behaviour even when $y = 0$. Plots with the other MCMC methods considered here, again did not suggest any problems. However comparisons with the chain autocorrelation of the approximate method showed superior performance for the latter, similar to that discussed in Section 7.

Tables 1–3 demonstrate that the risk properties of the AHB estimator are similar to those of the exact HB methods, verifying the remarkably accurate performance of our approximate Gibbs sampling approach. Note however, that in most cases the AHB estimator outperforms the exact algorithms; this can be explained by the better convergence properties of the approximate method, as discussed later in Section 7. The results also reveal that both algorithms using the approximation developed in Section 4 (i.e. the AHB and mixture-proposal HB), have lower risk than the more commonly used normal-proposal Metropolis-within-Gibbs scheme. This underlines the importance of employing a good proposal distribution in the chosen MCMC algorithm.

Fig. 2. (a) MCMC trace and (b) autocorrelation plots of the approximate Gibbs sampling method for a simulated data set generated with $\mathrm{E}(\theta_i) = 5$ and $\mathrm{var}(\theta_i) = 10$. The data values are: $y_1 = 14$, $y_2 = 9$, $y_3 = 11$, $y_4 = 0$, $y_5 = 2$, $y_6 = 14$, $y_7 = 4$, $y_8 = 0$, $y_9 = 12$, $y_{10} = 3$.

(a)



(b)

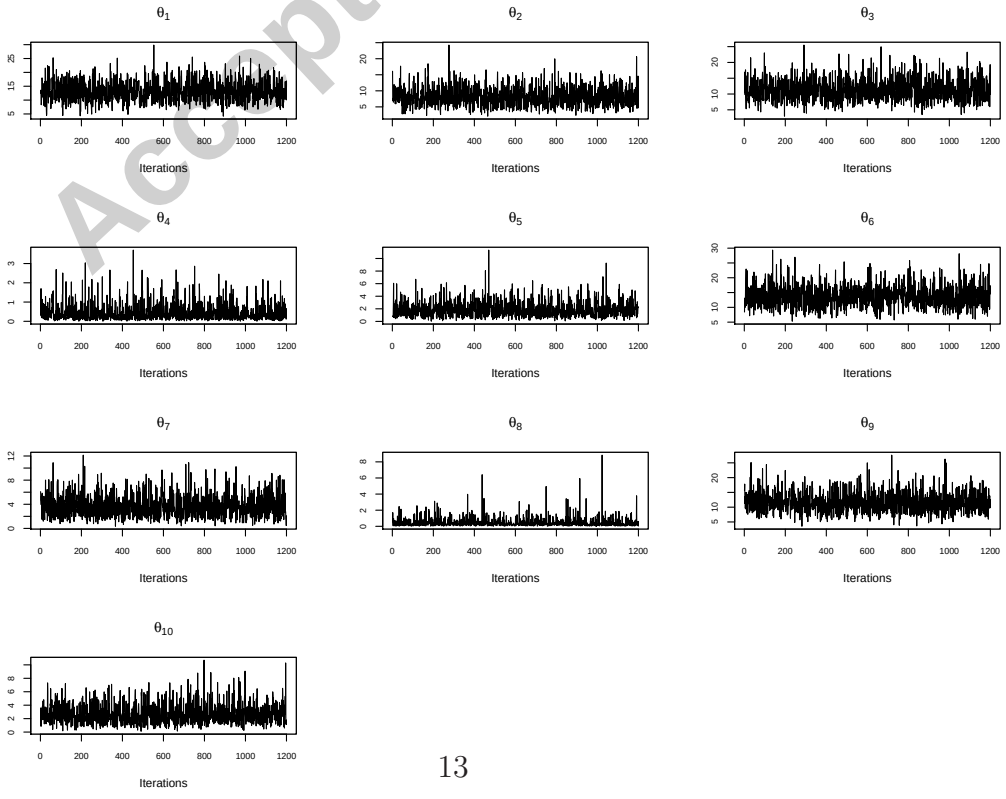


Table 1
Estimated Bayes risk (23) of the approximate hierarchical Bayes and the independence Metropolis–Hastings estimator of Section 5 under squared error loss (19), normalised squared error loss (20) and maximum component squared error loss (21). Number of observations is $m = 10$. The estimated risk of a hierarchical Bayes estimator using a normal proposal, the empirical Bayes estimator (22), the Morris modification and the MLE are also given. The number of simulations for the evaluation of (23) was $N = 10^4$; simulation standard errors are reported in brackets.

$\mathrm{E}(\theta_i)$	5.0		10.0	
$\mathrm{var}(\theta_i)$	2.5	10.0	5.0	20.0
Squared error loss				
Approximate HB	2.532 (.002)	3.689 (.002)	5.043 (.003)	7.443 (.004)
HB (mixture prop.)	2.549 (.002)	3.702 (.002)	5.074 (.003)	7.460 (.004)
HB (normal prop.)	2.704 (.005)	3.718 (.003)	6.796 (.044)	7.530 (.006)
Empirical Bayes	2.329 (.001)	3.784 (.003)	4.728 (.003)	7.669 (.004)
EB (Morris)	2.407 (.002)	3.700 (.002)	4.891 (.003)	7.509 (.004)
MLE	4.975 (.003)	4.955 (.003)	10.058 (.005)	9.936 (.005)
Normalised squared error loss				
Approximate HB	0.498 (.000)	0.723 (.000)	0.500 (.000)	0.737 (.000)
HB (mixture prop.)	0.501 (.000)	0.727 (.000)	0.504 (.000)	0.738 (.000)
HB (normal prop.)	0.536 (.001)	0.732 (.000)	0.678 (.004)	0.749 (.001)
Empirical Bayes	0.466 (.000)	0.781 (.000)	0.475 (.000)	0.777 (.000)
EB (Morris)	0.477 (.000)	0.741 (.000)	0.488 (.000)	0.750 (.000)
MLE	0.995 (.001)	0.992 (.001)	1.006 (.001)	0.995 (.001)
Max. component squared error loss				
Approximate HB	10.490 (.089)	17.114 (.170)	20.051 (.153)	31.855 (.259)
HB (mixture prop.)	10.542 (.090)	17.206 (.163)	20.085 (.148)	32.055 (.253)
HB (normal prop.)	11.193 (.169)	17.260 (.168)	25.125 (.872)	32.310 (.299)
Empirical Bayes	9.524 (.079)	17.485 (.170)	18.514 (.131)	32.892 (.353)
EB (Morris)	9.860 (.084)	17.037 (.163)	19.176 (.142)	32.044 (.254)
MLE	20.053 (.143)	21.889 (.186)	39.486 (.250)	41.195 (.294)

Table 1 shows that for $m = 10$ the AHB and the mixture-proposal HB estimators possess smaller risk than the EB methods when the true variance of $\theta_i, i = 1, \ldots, m$, is large, as expected due to the vague hyperpriors used. For $m = 30$, Table 2 demonstrates that as the number of observations increases the risk is greatly reduced, and the suggested estimators perform better than the EB methods for almost all examined true prior distributions and loss functions.

Table 2
Estimated Bayes risk (23) of the approximate hierarchical Bayes and the independence Metropolis–Hastings estimator of Section 5 under squared error loss (19), normalised squared error loss (20) and maximum component squared error loss (21). Number of observations is $m = 30$. The estimated risk of a hierarchical Bayes estimator using a normal proposal, the empirical Bayes estimator (22), the Morris modification and the MLE are also given. The number of simulations for the evaluation of (23) was $N = 10^4$; simulation standard errors are reported in brackets.

$\mathrm{E}(\theta_i)$	5.0		10.0	
$\mathrm{var}(\theta_i)$	2.5	10.0	5.0	20.0
Squared error loss				
Approximate HB	1.907 (.000)	3.431 (.000)	3.810 (.000)	6.954 (.001)
HB (mixture prop.)	1.948 (.000)	3.436 (.000)	3.864 (.000)	6.934 (.001)
HB (normal prop.)	1.959 (.000)	3.453 (.001)	3.886 (.000)	6.939 (.001)
Empirical Bayes	1.973 (.000)	3.499 (.001)	3.931 (.000)	7.014 (.001)
EB (Morris)	1.967 (.000)	3.483 (.001)	3.917 (.000)	6.991 (.001)
MLE	5.006 (.001)	4.998 (.001)	9.977 (.001)	10.022 (.001)
Normalised squared error loss				
Approximate HB	0.379 (.000)	0.686 (.000)	0.382 (.000)	0.694 (.000)
HB (mixture prop.)	0.387 (.000)	0.684 (.000)	0.386 (.000)	0.694 (.000)
HB (normal prop.)	0.389 (.000)	0.686 (.000)	0.387 (.000)	0.695 (.000)
Empirical Bayes	0.397 (.000)	0.710 (.000)	0.395 (.000)	0.709 (.000)
EB (Morris)	0.394 (.000)	0.703 (.000)	0.393 (.000)	0.704 (.000)
MLE	1.000 (.000)	0.997 (.000)	0.999 (.000)	1.003 (.000)
Max. component squared error loss				
Approximate HB	13.095 (.083)	28.397 (.218)	23.857 (.135)	49.812 (.309)
HB (mixture prop.)	13.292 (.084)	28.289 (.211)	24.090 (.132)	49.371 (.316)
HB (normal prop.)	13.478 (.087)	28.466 (.221)	24.518 (.138)	49.320 (.306)
Empirical Bayes	13.368 (.085)	28.852 (.223)	24.441 (.136)	49.954 (.330)
EB (Morris)	13.227 (.083)	28.345 (.216)	24.255 (.132)	49.297 (.320)
MLE	31.832 (.177)	36.282 (.235)	59.739 (.295)	65.836 (.366)

In Table 3 we investigate the robustness of our methods when the true θ_i values are generated from a gamma distribution with its parameters suitably chosen to match the selected combinations of the true mean and variance. The table reveals that the suggested methodology is robust under the assumption that the data come from a Poisson/gamma model, as the results are similar to those of Table 2. However, we notice that the large true variance no longer favours the HB methods as strongly as before, as the linear rule used for the EB estimators gives exactly the posterior mean under the conjugate formulation. Finally, the reduced risk of the Bayes methods under the maximum component

loss function (21) when Poisson/gamma data are considered, may be explained by the less likely presence of possible outliers under this assumption.

Table 3
Estimated Bayes risk (23) of the approximate hierarchical Bayes and the independence Metropolis–Hastings estimator of Section 5 under squared error loss (19), normalised squared error loss (20) and maximum component squared error loss (21), when data are generated from a Poisson/gamma model. Number of observations is $m = 30$. The estimated risk of a hierarchical Bayes estimator using a normal proposal, the empirical Bayes estimator (22), the Morris modification and the MLE are also given. The number of simulations for the evaluation of (23) was $N = 10^4$; simulation standard errors are reported in brackets.

$\mathrm{E}(\theta_i)$	5.0		10.0	
$\mathrm{var}(\theta_i)$	2.5	10.0	5.0	20.0
Squared error loss				
Approximate HB	1.948 (.000)	3.609 (.000)	3.876 (.000)	7.103 (.001)
HB (mixture prop.)	1.970 (.000)	3.589 (.000)	3.893 (.000)	7.098 (.001)
HB (normal prop.)	1.977 (.000)	3.614 (.000)	3.913 (.000)	7.146 (.001)
Empirical Bayes	1.976 (.000)	3.541 (.000)	3.939 (.000)	7.046 (.001)
EB (Morris)	1.972 (.000)	3.528 (.000)	3.928 (.000)	7.025 (.001)
MLE	5.022 (.001)	5.002 (.001)	10.007 (.001)	9.985 (.001)
Normalised squared error loss				
Approximate HB	0.408 (.000)	0.883 (.000)	0.396 (.000)	0.760 (.000)
HB (mixture prop.)	0.408 (.000)	0.858 (.000)	0.396 (.000)	0.762 (.000)
HB (normal prop.)	0.409 (.000)	0.865 (.000)	0.398 (.000)	0.763 (.000)
Empirical Bayes	0.419 (.000)	0.869 (.000)	0.406 (.000)	0.766 (.000)
EB (Morris)	0.414 (.000)	0.843 (.000)	0.403 (.000)	0.755 (.000)
MLE	1.004 (.000)	1.001 (.000)	1.001 (.000)	1.001 (.000)
Max. component squared error loss				
Approximate HB	12.520 (.075)	27.928 (.184)	22.906 (.116)	47.789 (.277)
HB (mixture prop.)	12.599 (.076)	27.517 (.181)	23.254 (.120)	47.619 (.275)
HB (normal prop.)	12.846 (.077)	27.895 (.178)	23.475 (.122)	48.990 (.286)
Empirical Bayes	12.359 (.071)	26.774 (.180)	23.222 (.116)	46.801 (.275)
EB (Morris)	12.346 (.072)	26.585 (.177)	23.186 (.116)	46.549 (.270)
MLE	31.772 (.178)	36.896 (.231)	59.557 (.287)	65.486 (.356)

7 Application to insurance data

We apply the approach outlined in Sections 3–5 to data on Permanent Health Insurance claim inceptions for 1999. The data set that we analyse concerns inceptions based on individual policies (of a deferred period of 26 weeks) for a female population of all occupational classes, grouped by age. The data are published by the Continuous Mortality Investigation Bureau (C.M.I.B., 1999) and presented here in Table 4. The number of observed claims, Y_i, is given for each of 9 age groups (covering a population spanning from 18 to 64 years of age), together with the total time of exposure (in years), E_i, per group. Column 5 of the table contains the number of expected claim inceptions calculated by the Continuous Mortality Investigation Bureau on the basis of estimates of sickness intensities (σ_i) and recovery intensities (ρ_i) using the Male Standard Experience for individual policies for 1975–1978 (C.M.I.B., 1991). In that previous analysis the estimation of the parameters of interest was based on a normal approximation $Y_i \dot{\sim} N(E_i\rho_i\sigma_i, V_iE_i\rho_i\sigma_i), i = 1, \ldots, m$, of the Poisson distribution of the number of claims. The factor V_i was introduced in the variance to account for overdispersion due to duplicate policies. The sickness intensities σ_i were modelled as an exponential polynomial of x_i (the midpoint of the age of group i), and an iterative generalised linear model procedure was used for the estimation.

Here we consider the claim intensities $\theta_i = \rho_i\sigma_i$, under the full hierarchical Bayesian model in (1)–(4) with $\mathbf{b} = (b_0, b_1, b_2)^{\mathrm{T}}$ and $\mathbf{x}_i^{\mathrm{T}} = (1, x_i, x_i^2)$, $i = 1, \ldots, 9$, assuming a vague prior distribution for the hyperparameters $\mathbf{b}$ and σ^2. The parameter θ_i is now regarded as the probability of an individual claim in age group i, and its logarithm is modelled as a quadratic function of age as previously suggested in C.M.I.B. (1991), i.e.

$$\log(\theta_i) = b_0 + b_1x_i + b_2x_i^2 + \varepsilon_i,\ i = 1, \ldots, 9.$$

Applying the methods of Sections 3–5, our approximate analysis produced the posterior estimates presented in Table 4. The posterior estimates of the claim intensities were virtually identical to estimates obtained using the exact mixture-proposal algorithm, as demonstrated in Fig. 3. The estimates of the regression coefficients b_0, b_1, b_2 and that of the variance component σ^2 are given in Table 5. The posterior estimates of b_2 suggest that the quadratic term may not be required when modelling $\log(\theta_i)$, and therefore we also considered a linear function of age for $\log(\theta_i)$. However, as Fig. 4 shows, the estimates of the claim intensities $\theta_1, \ldots, \theta_9$, do not change greatly under the linear modelling. The largest difference is observed in the estimate of θ_9, which is expected under the limited flexibility of the linear model to deal with $y_9 = 0$ following an increasing trend of observed values, and due to the high variability caused by

Table 4
Insurance claims based on individual policies (of a deferred period of 26 weeks) for a female population of all occupational classes (C.M.I.B., 1999), grouped by age. The third column of the table contains the total time of exposure in years for each of the 9 age groups; the fourth column gives the number of actual (observed) claims per group; the fifth column shows estimates of expected number of claims obtained by the C.M.I.B. based on the Male Standard Experience for individual policies for 1975–1978; posterior summaries of expected claims using our hierarchical model are presented in columns six and seven of the table; the last two columns give posterior summaries of the claim intensities $\theta_1, \ldots, \theta_9$. In the Bayesian analysis $\log(\theta_i)$ was modelled as a quadratic function of age.

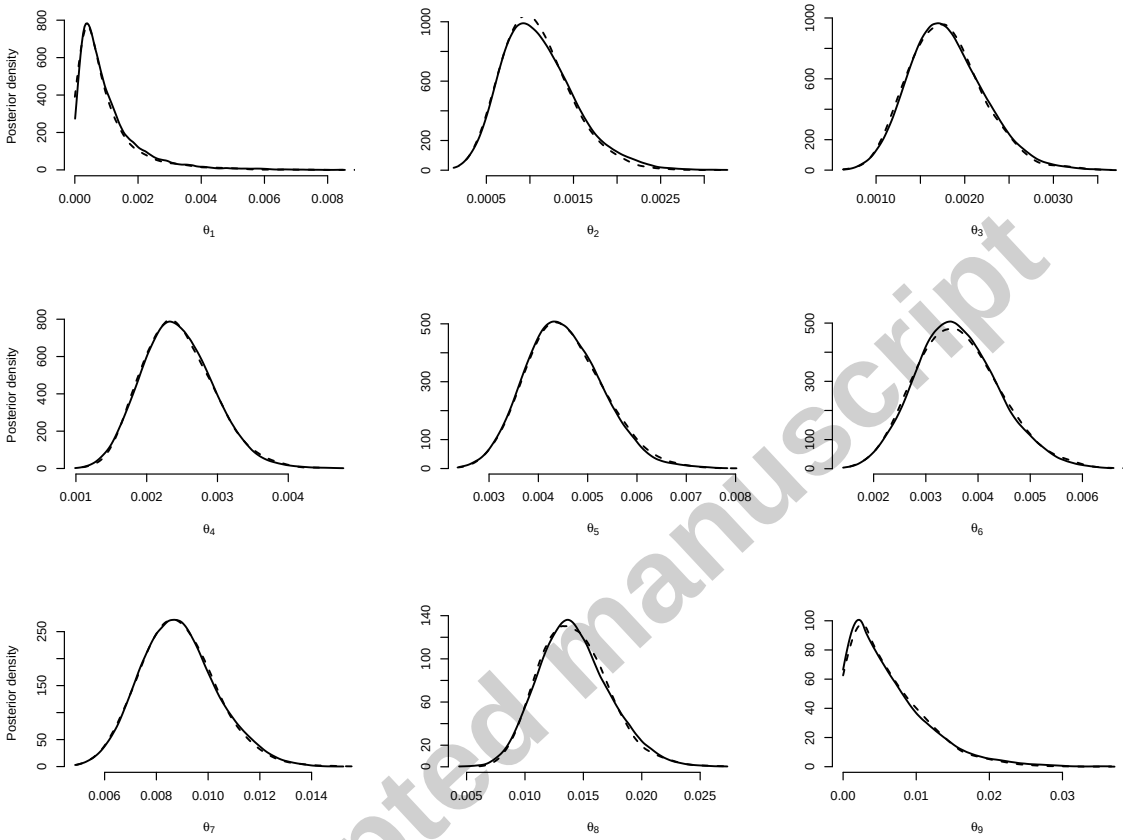
				Expected claims			Claim intensity	
Group	Age	Expos.	Claims		Posterior		Posterior	
i	group	E_i	y_i	C.M.I.B	mean	sd	mean	sd
1	18–24	646	1.0	0.5	0.6	1.0	0.0009	0.0010
2	25–29	5665	6.0	5.4	6.1	3.2	0.0011	0.0004
3	30–34	9472	17.0	12.5	16.8	5.6	0.0018	0.0004
4	35–39	8784	21.0	17.5	21.5	6.4	0.0024	0.0005
5	40–44	7176	33.0	22.8	32.4	8.0	0.0045	0.0008
6	45–49	5959	20.0	31.5	21.7	6.6	0.0036	0.0008
7	50–54	4070	37.0	36.9	35.7	8.3	0.0088	0.0014
8	55–59	1635	25.0	26.5	23.3	6.9	0.0142	0.0030
9	60–64	217	0.0	6.4	1.3	1.6	0.0058	0.0049

Table 5
Hierarchical Bayes estimates of the log-normal hyperparameters for the insurance data, when the logarithm of the claim intensity is modelled as a quadratic polynomial of age.

	b_0	b_1	b_2	σ^2
mean	−5.811	0.214	−0.002	1.293
sd	0.457	0.259	0.003	3.773
2.5%	−6.910	−0.227	−0.009	0.030
median	−5.758	0.184	−0.002	0.461
97.5%	−5.100	0.841	0.003	7.195

the low exposure. The Bayesian analysis produced smoothed estimates for the claim intensities and expected numbers of claims, allowing a moderate shrinkage effect as expected with the use of vague prior information in the model. Fig. 4 shows 95% credible intervals of the true claim intensities, revealing how the probability of a claim varies according to age group.

Fig. 3. Estimated posterior densities of the claim intensities $\theta_1, \ldots, \theta_9$, for the insurance data under the approximate Gibbs sampling approach (solid line) and under the exact independence Metropolis–Hastings algorithm using the mixture proposal (dashed line).



Estimation was based on 40000 sampled values, following a burn-in period of 5000 iterations, and the convergence of the algorithm was assessed and verified using various diagnostic criteria (Cowles and Carlin, 1996). We compare the convergence performance of the approximate Gibbs sampler and the exact independence Metropolis–Hastings approach described in Section 5 to that of the normal-proposal Metropolis-within-Gibbs method introduced in Section 6, and we also consider the widely used Gibbs sampler of *WinBUGS*. For the comparisons we employ the effective sample size (ESS; Brooks et al., 2003) and the Monte Carlo (MC) error of the four methods. ESS is a measure of sample size adjusted for the autocorrelation of the chain and provides, for a given parameter, the number of independent sampled values which corresponds to the number of dependent values produced by the Markov chain. For each method we consider the minimum ESS and the average MC error among model parameters. These are shown in Fig. 5 for a range of MCMC iterations, while the MC error for all θ_i parameters is given in Table 6. Both plots and the table demonstrate that the approximate Gibbs sampling approach mixes more

Fig. 4. 95% posterior credible intervals for the claim intensities $\theta_1, \ldots, \theta_9$, in the insurance data application. Solid and dashed lines correspond to modelling $\log(\theta_i)$ through a quadratic and a linear function of age respectively. The circles indicate the posterior means.

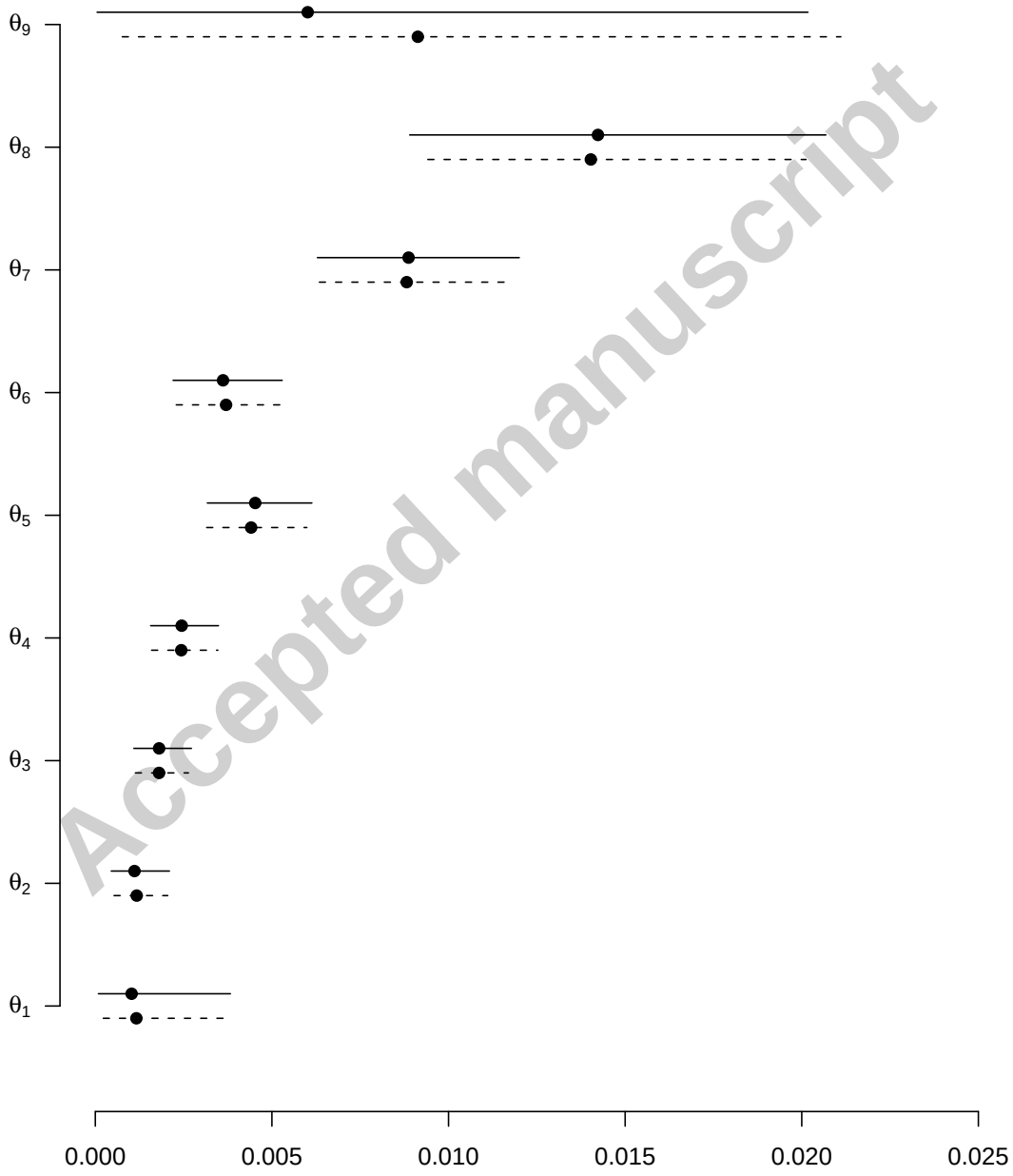
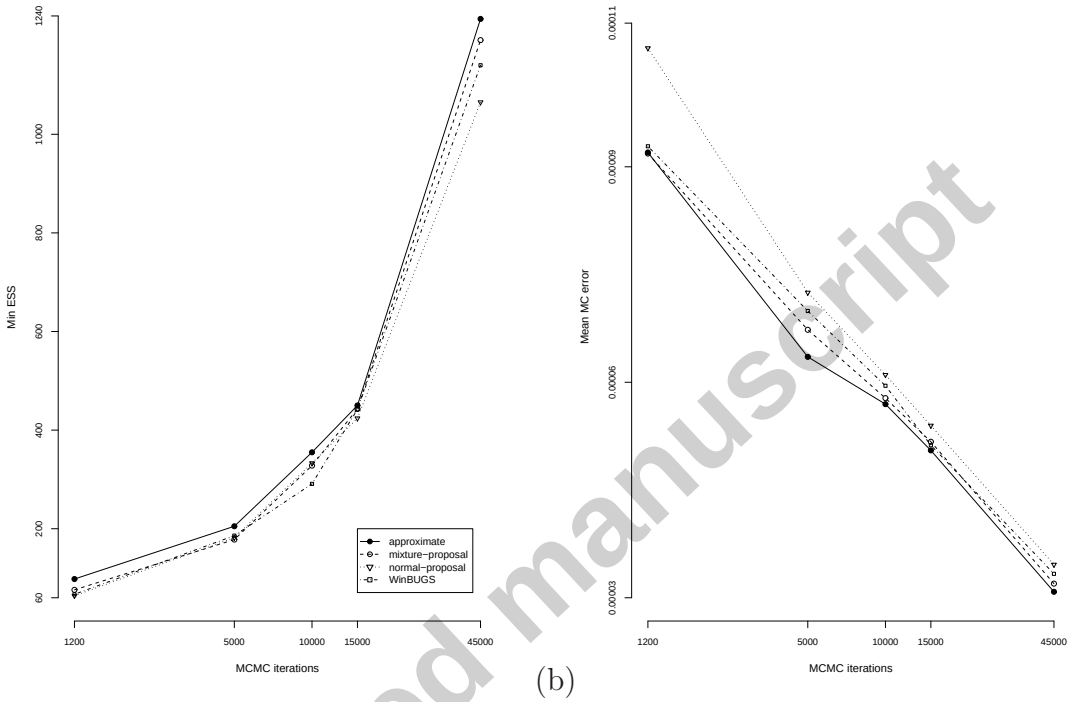


Fig. 5. (a) Minimum effective sample size (ESS) and (b) mean Monte Carlo error under a range of MCMC iterations for the insurance data, using: the approximate Gibbs sampling approach (solid line); the mixture-proposal independence Metropolis–Hastings chain (dashed line); a normal-proposal Metropolis-within-Gibbs algorithm (dotted line); and *WinBUGS* (dot-dashed line). A logarithmic scale has been used on the x axis.



(a) (b)

efficiently than the exact algorithms, while providing almost identical posterior estimates (as illustrated in Fig. 3). Chain autocorrelation plots for some of the parameters (with θ_9 being the slowest mixing Poisson rate) are shown in Fig. 6, demonstrating that the methods relying on the suggested mixture approximation are more efficient than the algorithm using the normal proposal distribution, while the approximate algorithm also outperforms *WinBUGS*. The acceptance rate of the mixture-proposal algorithm was 0.98 (average over all θ parameters), as compared to 0.87 for the normal-proposal method. We note in addition that, as expected, the approximate method is faster in terms of computer running time, requiring 4 seconds for 10000 iterations as compared to 4.75 seconds for the exact Metropolis–Hastings algorithms (19% relative increase in speed). The equivalent analysis in the *WinBUGS* software package required 5 seconds (approximate method 25% faster).

Table 6
Monte Carlo standard error ($\times 10^4$) of the posterior means of $\theta_1, \ldots, \theta_9$ for the insurance claims data set (with 45000 MCMC updates).

	Approximate	Mixture proposal	Normal proposal	*WinBUGS*
θ_1	0.2999	0.2621	0.3038	0.2757
θ_2	0.0645	0.0560	0.0709	0.0583
θ_3	0.0243	0.0221	0.0303	0.0230
θ_4	0.0318	0.0325	0.0498	0.0322
θ_5	0.0736	0.0797	0.0829	0.0816
θ_6	0.0785	0.0819	0.0941	0.0857
θ_7	0.0954	0.1020	0.1093	0.0982
θ_8	0.2153	0.2384	0.2722	0.2498
θ_9	1.8914	2.0024	2.1028	2.0956

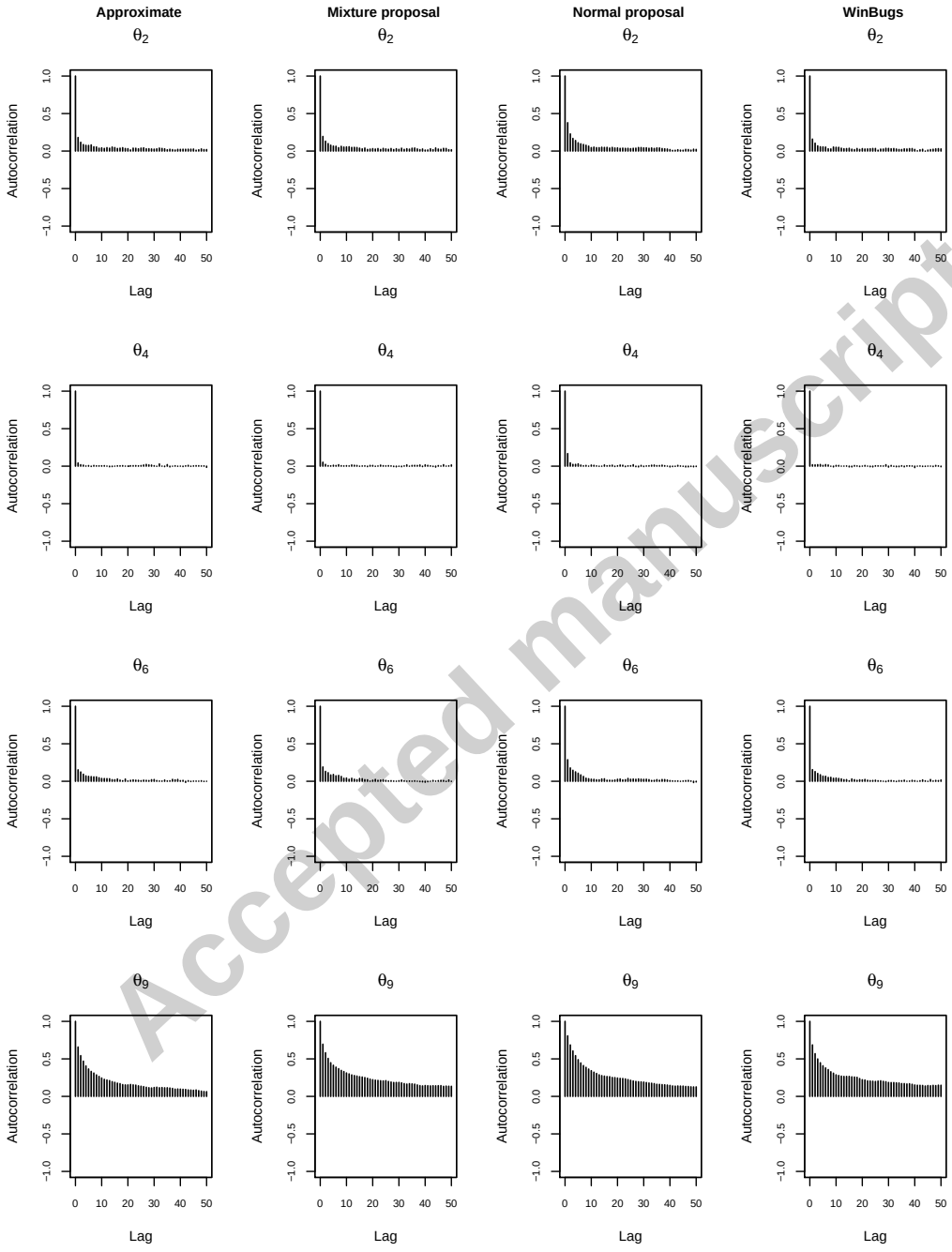
8 Discussion

We have presented an efficient approach for a fully Bayesian analysis of count data considering a Poisson/log-normal model. The log-normal prior distribution leads to a more dispersed marginal distribution for the data, when compared to other prior assumptions. The assessment of various Poisson models used to account for overdispersion in actuarial data is an active topic of current research (e.g. Ntzoufras et al., 2005). Employing a conjugate gamma prior provides some mathematical tractability through the linearity in the conditional posterior moments (e.g. Christiansen and Morris, 1997). However this is not of critical importance when the interest is on the entire posterior distribution under a fully Bayesian formulation.

Simulation from the conditional posterior distribution of θ_i using a log-normal or a gamma distribution as the basis for the approximation (instead of the mixture in Section 4) did not provide the level of accuracy that is required in the approximate method, whereas employing a matched normal proposal in a Metropolis-within-Gibbs scheme was less efficient. Furthermore, none of these schemes add any substantial computational efficiency to our approach, as most of the computational effort in the methods is required for the calculation of the moments $\mathrm{E}(\theta_i|\boldsymbol{b}, \sigma^2, \mathbf{y})$. As described in Section 4, matching these with the moments of a log-normal/gamma mixture is computationally easy, and so is simulating from such a mixture.

It is worth noting that using the approximations described in Section 4.1, our methods perform remarkably well also in cases where the prior variance is large relative to a small prior mean (e.g. when $\mathrm{E}(\theta_i) = 5$, $\mathrm{var}(\theta_i) = 10$ in Tables 1–3). This would be the typical situation where the marginal distribution of the

Fig. 6. MCMC autocorrelation plots for $\theta_2, \theta_4, \theta_6$ and θ_9 in the insurance data application, using: the approximate Gibbs sampling approach (first panel from left); the mixture-proposal independence Metropolis–Hastings chain (second panel); the normal-proposal Metropolis-within-Gibbs algorithm (third panel); and *WinBUGS* (fourth panel).



data would accommodate a moderate number of zero values in the observed sample. Analysis of such real and simulated data sets, also confirmed the efficient performance of the proposed algorithms.

Regarding potential extensions of our results in related hierarchical models of different distributional structure, we note that expression (14) holds irrespective of the form of the prior distribution $\pi(\theta_i)$. This implies that our methodology can also be applied to cases where the distribution of the Poisson rates is given by other parametric families, once a suitable density to provide the basis for the approximation of $p(\theta_i|\boldsymbol{b}, \sigma^2, \mathbf{y})$ is identified. For example, in the case of the usual conjugate gamma prior, the moments in (14) are easily available noticing that $p(y_i)$ is a negative binomial density; however, no approximation is required in this case as the conditional moments of θ_i and the entire conditional density are analytically available. Nevertheless, our work can be extended to allow for prior uncertainty described by distributions taken, for example, from the Pareto or Weibull families.

In general, Bayesian methods provide estimators that are based on a given set of observed data; however, it is also desirable that they display a robust and repeatable performance. The simulation study showed that our hierarchical Bayes estimate of the posterior mean exhibits good risk properties in various scenarios regarding the true prior distribution and loss function of interest. Combining the Bayesian methodology with frequentist criteria of repeatability performance evaluation can be useful when we wish to derive good inferential procedures irrespective of the underlying philosophical perspective (e.g. Samaniego and Reneau, 1994; Carlin and Louis, 2000).

Acknowledgements

The authors would like to thank Professor Tom Leonard for many helpful discussions and suggestions. We are grateful to Professor Howard R. Waters for discussions concerning the data on insurance claims. We would also like to thank the associate editor and two anonymous reviewers for their insightful and constructive comments on an earlier version of the manuscript that greatly improved the content and presentation of this paper.

Appendix A: Minimisation of the entropy distance $\mathcal{D}I(\alpha_i, \omega_i^2)$

Working in a similar way as in (14), we notice that for any real number t

$$p(\lambda_i|y_i) \propto L(\lambda_i|y_i)\ \pi(\lambda_i) = e^{-t\lambda_i} L(\lambda_i|y_i + t)\ \pi(\lambda_i), \tag{24}$$

where

$$L(\lambda_i|y_i + t) = \frac{E_i^{y_i+t} \exp\left\{\lambda_i(y_i + t) - E_i e^{\lambda_i}\right\}}{\Gamma(y_i + t)},$$

gives the probability density function of the random variable $\lambda_i = \log(\theta_i)$, with $\theta_i \sim \text{Ga}(y_i + t, E_i)$. Hence, noting that the logarithm of a gamma variable is approximately normally distributed (Gelman et al., 2004, p.579; Johnson et al., 1994, p.383) we approximate $L(\lambda_i|y_i+t)$ with the probability density function of a normal $N(\delta_i, v_i^2)$ distribution, denoted by $L^*(\lambda_i|y_i + t)$. We consider the entropy distance (O'Hagan and Forster, 2004)

$$\mathcal{D}I(\delta_i, v_i^2) = \text{E}\left\{\log \frac{L^*(\lambda_i|y_i + t)}{L(\lambda_i|y_i + t)}\right\},$$

where the expectation corresponds to the $N(\delta_i, v_i^2)$ distribution for λ_i. This will give

$$\mathcal{D}I(\delta_i, v_i^2) = -\frac{1}{2}\log(2\pi) + \log\{\Gamma(y_i + t)\} - (y_i + t)\log(E_i) \\ - \log v_i - \frac{1}{2} - \delta_i(y_i + t) + E_i e^{\delta_i + \frac{1}{2}v_i^2},$$

and setting the first derivatives with respect to δ_i, v_i^2 equal to zero, the entropy distance is minimised for

$$\delta_i = \log\left(\frac{y_i + t}{E_i}\right) - \frac{1}{2}(y_i + t)^{-1}, \quad v_i^2 = (y_i + t)^{-1}.$$

Then taking into account the multiplicative factor $e^{-t\lambda_i}$ in (24) we obtain a normal approximation to $e^{-t\lambda_i}L(\lambda_i|y_i+t)$ with mean $l_i = \delta_i - tv_i^2$ and variance v_i^2. Setting $t = \frac{1}{2}$, which is the usual bias correction (Plackett, 1974, p.3), we can write

$$l_i = \log\left(\frac{y_i + \frac{1}{2}}{E_i}\right) - \left(y_i + \frac{1}{2}\right)^{-1}, \qquad v_i^2 = \left(y_i + \frac{1}{2}\right)^{-1}.$$

The conjugacy of the normal $N(l_i, v_i^2)$ likelihood with the normal $N(\mathbf{x}_i^{\text{T}}\mathbf{b}, \sigma^2)$ prior $\pi(\lambda_i)$ implies that, $p^*(\lambda_i|y_i)$ is a $N(\tilde{\alpha}_i, \tilde{\omega}_i^2)$ density with

$$\tilde{\alpha}_i = \frac{v_i^{-2}l_i + \sigma^{-2}\mathbf{x}_i^{\text{T}}\mathbf{b}}{v_i^{-2} + \sigma^{-2}}, \qquad \tilde{\omega}_i^2 = \left(v_i^{-2} + \sigma^{-2}\right)^{-1}. \tag{25}$$

Appendix B: Discrete approximation to the standard normal distribution

We derive an l-point discrete approximation to the standard normal distribution by matching the first 10 moments of the exact and the approximating distribution. For the normal $N(0, 1)$ distribution the moments of odd order are

equal to zero, while those of even order are provided by the following equation (Stuart and Ord 1994):

$$\mathrm{E}\left(Z^{2r}\right) = \frac{(2r)!}{2^r r!}, \qquad r = 1, \ldots, 5.$$

For a symmetric distribution of a discrete random variable X evaluated at the points $\gamma_1, \gamma_2, \ldots, \gamma_l$, with probabilities $p_1, p_2, \ldots, p_l$, and if we take the points γ_j to be equally spaced on a suitably selected grid with the distance between two successive points equal to a fixed value γ, the corresponding moments of even order are given by

$$\mathrm{E}\left(X^{2r}\right) = 2 \sum_{j=0}^{\frac{l-1}{2}} (j\gamma)^{2r} p_j, \qquad r = 1, \ldots, 5.$$

Then by solving the system of equations

$$\frac{(2r)!}{2^r r!} = 2 \sum_{j=0}^{\frac{l-1}{2}} (j\gamma)^{2r} p_j, \qquad r = 1, \ldots, 5,$$

we obtain the probabilities $p_1, p_2, \ldots, p_{\frac{l-1}{2}}$ and $p_0 = 1 - 2\sum_{j=1}^{\frac{l-1}{2}} p_j$, which correspond to the points lying on the non-negative part of the x-axis. Clearly, the same probabilities correspond to the equivalent points on the negative axis.

References

Ainsworth, L.M., Dean, C.B., 2006. Approximate inference for disease mapping. Comput. Statist. Data Anal. 50, 2552–2570.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York.

Brooks, S.P., Giudici, P., Roberts, G.O., 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions (with discussion). J.R. Statist. Soc. B 65, 3–55.

Candel, M.J.J.M., 2007. Empirical Bayes estimators of the random intercept in multilevel analysis: performance of the classical, Morris and Rao version. Comput. Statist. Data Anal. 51, 3027–3040.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis. second ed. Chapman & Hall/CRC, Boca Raton.

Christiansen, C.L., Morris, C.N., 1997. Hierarchical Poisson regression modeling. J. Amer. Statist. Assoc. 92, 618–632.

Clayton, D., Kaldor, J., 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671–681.

C.M.I.B., 1991. CMI Report No. 12. The Faculty of Actuaries and the Institute of Actuaries, United Kingdom.

C.M.I.B., 1999. Analysis of IP Claim Inceptions and Claim Terminations by Occupational Class. The Institute of Actuaries and the Faculty of Actuaries, United Kingdom.

Cowles, M.K., Carlin, B.P., 1996. Markov chain Monte Carlo convergence diagnostics: a comparative review. J. Amer. Statist. Assoc. 91, 883–904.

Currie, I.D., Waters, H.R., 1991. On modelling select mortality. J. Instit. Actuar. 118, 453–481.

Czado, C., Delwarde, A., Denuit, M., 2005. Bayesian Poisson log-bilinear mortality projections. Insur.: Math. Econ. 36, 260–284.

Damien, P., Wakefield, J., Walker, S., 1999. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. J. Roy. Statist. Soc. B 61, 331–344.

Efron, B., Morris, C., 1973. Stein's estimation rule and its competitors—An empirical Bayes approach. J. Amer. Statist. Assoc. 68, 117–130.

Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85, 398–409.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. second ed. Chapman and Hall/CRC, London.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences (with discussion). Statist. Sci. 7, 457–511.

George, E.I., Makov, U.E., Smith, A.F.M., 1994. Fully Bayesian hierarchical analysis for exponential families via Monte Carlo computation. In: Freeman P.R. and Smith A.F.M. (Eds.), Aspects of Uncertainty: A Tribute to D.V. Lindley. Wiley, Chichester, pp. 181–199.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57, 1317–1339.

Grandell, J., 1997. Mixed Poisson Processes. Chapman & Hall/CRC, New York.

Haastrup, S., 2000. Comparison of some Bayesian analyses of heterogeneity in group life insurance. Scand. Actuarial J. 1, 2–16.

Haberman, S., Renshaw, A.E., 1996. Generalized linear models in actuarial science. The Statistician 45, 407–436.

Hudson, H.M., 1978. A natural identity for exponential families with applications in multiparameter estimation. Ann. Statist. 6, 473–484.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. Continuous Univariate Distributions, vol. 1. second ed. Wiley, New York.

Leonard, T., Hsu, J.S.J., 1999. Bayesian Methods: An Analysis for Statisticians and Interdisciplinary Researchers. Cambridge University Press, Cambridge.

Makov, U.E., Smith, A.F.M., Liu, Y.-H., 1996. Bayesian methods in actuarial science. The Statistician 45, 503–515.

Morris, C.N., 1983. Discussion of 'Construction of improved estimators in multiparameter estimation for discrete exponential families', by Ghosh M., Hwang J.T., and Tsui K.W. Ann. Statist. 11, 372–374.

Ntzoufras, I., Katsis, A., Karlis, D., 2005. Bayesian assessment of the dis-

tribution of insurance claim counts using reversible jump MCMC. North. Amer. Actuar. J. 9(3), 90–108.

O'Hagan, A., Forster, J., 2004. Kendall's Advanced Theory of Statistics, vol. 2B, Bayesian Inference. second ed. Arnold, London.

Plackett, R.L., 1974. The Analysis of Categorical Data. London: Griffin.

Samaniego, F.J., Reneau, D.M., 1994. Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. J. Amer. Statist. Assoc. 89, 947–957.

Spiegelhalter, D.J., Thomas, A., Best, N.G., Gilks, W.R., 1996. BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge.

Stuart, A., Ord, J.K., 1994. Kendall's Advanced Theory of Statistics, vol. 1, Distribution Theory. sixth ed. Edward Arnold, London.

Tierney, L., 1994. Markov chains for exploring posterior distributions (with discussion). Ann. Statist. 22, 1701–1762.

Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, New York.