BAYESIAN METHODS FOR POISSON MODELS

George Streftaris

Doctor of Philosophy
University of Edinburgh
2000

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*George Streftaris*)

Acknowledgements

I would like to thank my supervisor, Dr. Bruce Worton, for his guidance, help and constant encouragement and support throughout this work. I am also grateful to Professor Tom Leonard for many helpful discussions and suggestions, while acting as my second supervisor.

I would also like to express my gratitude to the State Scholarships Foundation of Greece, for funding this research.

Abstract

To account for overdispersion in count data, that is variation in excess of that justified from the assumed model, one may consider an additional source of variation, by assuming that each observation, Y_i, $i = 1, \ldots, m$, arises from a conditionally independent Poisson distribution, given its respective mean θ_i, $i = 1, \ldots, m$.

We review various frequentist methods for the estimation of the Poisson parameters θ_i, $i = 1, \ldots, m$, which are based on the inadmissibility of the usual unbiased maximum likelihood estimator, in terms of the associated risk in dimensions greater than two. The so called shrinkage estimators adjust the maximum likelihood estimates towards a fixed or data-determined point, abandoning unbiasedness in favour of lower risk.

Inferences for the parameters of interest can also be drawn employing Bayesian methods. Conjugate models are often adopted to facilitate the computational procedure. In this thesis we assume a nonconjugate log-normal prior distribution, which allows for more dispersion in the Poisson means and can also accommodate a correlation structure. We derive two empirical Bayes estimators, which approximate the posterior mean. The first is based on a linear shrinkage rule, while the second employs a non-iterative importance sampling technique. The frequency properties of the two estimators in terms of average risk are assessed and compared to other estimating approaches proposed in the literature.

A full hierarchical Bayes analysis is also considered, assuming both informative and vague prior distributions at the lower stage of the hierarchy. Some analytical posterior inferences, based on simple approximations are obtained. We then employ stochastic simulation techniques, suggesting two Markov chain Monte Carlo methods which involve the Gibbs sampler and a hybrid strategy. They rely on a log-normal/gamma mixture approximation to the full conditional posterior distribution of the parameters θ_i, $i = 1, \ldots, m$. The shrinkage behaviour of the hierarchical Bayes estimator is explored, and its average risk is examined through frequency simulations. Examples and applications of the considered methods are given throughout the thesis.

Table of Contents

Chapter 1

Introduction

1.1 Estimation of several Poisson means

Poisson models are widely used to describe the distribution of events occurring independently throughout time or space. In many situations the numbers of events, expressed as count data, may exhibit variation which exceeds what would be justified under the assumed model. To account for this overdispersion, it is common to assume an additional source of variation in the data, by considering that given a parameter θ_i, each observation Y_i follows a conditionally independent Poisson distribution with mean θ_i, $i = 1, \ldots, m$, i.e.

$$Y_i|\theta_i \overset{ind}{\sim} \text{Poisson}(\theta_i) \qquad i = 1, \ldots, m, \tag{1.1}$$

where $\theta_i > 0$. In doing so, not only can we accommodate some of the extra-Poisson variation, but we may also exploit the information included in all m parameters to obtain better estimates for the individual means θ_i, $i = 1, \ldots, m$.

The main objective of this thesis is to develop methods for deriving estimators of the Poisson parameters θ_i, $i = 1, \ldots, m$, in a Bayesian framework, and investigate their frequency properties under the criterion of the associated average risk. The latter is a measure of the discrepancy between the produced estimates and the true parameters, averaged over both the conditional distribution of the data and the prior distribution, and will be formally introduced in Chapter 3. Combining the Bayesian methodology with frequentist criteria of performance evaluation, can be useful when one wishes to derive good inferential procedures irrespective of the underlying philosophical perspective (e.g. see Carlin and Louis, 1996). The estimation problem is approached from both the empirical and the hierarchical Bayesian point of view, employing analytical approximations and Monte Carlo integration methods. We also wish to explore the behaviour of the estimators when information from all the m simultaneously considered parameters is combined.

1.2 Unbiased and shrinkage estimation for the Poisson means

To draw inferences about the Poisson means θ_i, $i = 1, \ldots, m$, one may employ the usual maximum likelihood (ML) method. By using the Poisson probability function given in Appendix A, the log-likelihood function of θ_i, denoted as $l(\theta_i|y_i)$, is given by

$$l(\theta_i|y_i) = y_i \log(\theta_i) - \theta_i - \log(y_i!), \qquad i = 1, \ldots, m, \qquad \theta_i > 0.$$

Straightforward maximisation of this log-likelihood function implies that the maximum likelihood estimator (MLE) of the Poisson means θ_i is simply given by

$$\hat{\theta}_i^{\text{MLE}} = Y_i, \qquad i = 1, \ldots, m.$$

This can be shown to be the uniformly minimum variance unbiased estimator (UMVUE) for θ_i. However, in estimating θ_i by y_i, one ignores the remaining components of the data vector $\mathbf{y} = (y_1, y_2, \ldots, y_m)^{\mathrm{T}}$, which can be important in situations where the estimation of each individual element of the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\mathrm{T}}$ may benefit from the information incorporated in the entire parameter vector. For instance, one may utilise the occurrence of a disease in many neighbouring geographical regions, to obtain better estimates of the number of incidents of the disease in a single small area (e.g. see Clayton and Kaldor, 1987). In fact, as discussed in Chapter 2, the MLE is inadmissible under various loss functions when two or more conditionally independent Poisson distributions are involved. This means that there exists at least one estimator with smaller risk than the MLE. It can be shown that estimating methods which shrink the ML estimates y_i, $i = 1, \ldots, m$, towards a point that often depends on some or all of the remaining components of the data vector $\mathbf{y}$, can result in estimators with better precision, and thus superior risk properties. These so-called shrinkage estimators are no longer unbiased for θ_i, $i = 1, \ldots, m$, but they have smaller variance than the MLE. The mean squared error (MSE) of an estimator, i.e. its risk under a squared error loss function, which can be expressed as the sum of the variance and the bias squared, motivates the intuition behind the trade-off between these two characteristics of the estimator, in order to obtain estimates that are closer to the true value of the parameters of interest. We therefore wish to estimate the Poisson means θ_i, $i = 1, \ldots, m$, in such a manner that we can exploit the multiparameter nature of the problem, to derive better inferences.

1.3 Bayesian analysis

According to the Bayesian approach, a prior structure is assumed for the parameters $\theta_i,\ i = 1, \ldots, m$, in addition to the distributional assumption in (1.1). Bayesian methods naturally behave as shrinkage estimators, exploiting the relation between $\theta_i,\ i = 1, \ldots, m$, as this is provided through the prior distribution. Inferences for the parameters of interest are based on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, or the appropriate marginal distributions, which is derived by suitably applying Bayes' theorem. The latter gives

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\boldsymbol{\theta}|\mathbf{y})\ \pi(\boldsymbol{\theta})}{f(\mathbf{y})}, \tag{1.2}$$

where $L(\boldsymbol{\theta}|\mathbf{y})$ and $\pi(\boldsymbol{\theta})$ denote the likelihood and the prior distribution of the parameter vector $\boldsymbol{\theta}$ respectively, and $f(\mathbf{y})$ is the marginal density of the data.

The form of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ depends on the prior assumptions for the parameter vector $\boldsymbol{\theta}$. Conjugate prior structures are often used, since they offer an analytically closed form of the posterior distribution of interest. For the Poisson distribution conjugacy is achieved using a gamma prior. We let $\text{Ga}(a, b)$ denote the gamma distribution with mean $\frac{a}{b}$ and variance $\frac{a}{b^2}$, where $a, b > 0$. The probability density function of this distribution is given in Appendix A. If we assume that the Poisson means $\theta_i,\ i = 1, \ldots, m$, independently follow a $\text{Ga}(a, b)$ distribution, then the conjugate Bayesian formulation can be expressed as

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \theta_i|a, b &\overset{iid}{\sim} \text{Ga}(a, b), \end{aligned} \tag{1.3}$$

for $i = 1, \ldots, m$. One advantage of using this Poisson/gamma conjugate structure is the mathematical convenience in the derivation of the posterior distributions of the parameters $\theta_i,\ i = 1, \ldots, m$. Bayes' formula in (1.2) implies that

$$\begin{aligned} p(\theta_i|\mathbf{y}) &\propto L(\theta_i|y_i)\ \pi(\theta_i) \\ &\propto \theta_i^{y_i} e^{-\theta_i}\ \theta_i^{a-1} e^{-b\theta_i} \\ &= \theta_i^{a+y_i-1} e^{-(b+1)\theta_i}, \end{aligned}$$

meaning that the posterior distribution of $\theta_i,\ i = 1, \ldots, m$, is given as an updated $\text{Ga}(a + y_i, b + 1)$ distribution. We can then use the latter to obtain posterior inferences for the Poisson means $\theta_i,\ i = 1, \ldots, m$.

1.3.1 The Poisson/log-normal formulation

In this thesis we focus on an alternative Bayesian structure. We assume that the Poisson parameters $\theta_i,\ i = 1, \ldots, m$, are independently and identically distributed according to a log-normal distribution with parameters μ and σ^2. If we

let the notation $\mathrm{LN}(\mu, \sigma^2)$ indicate such a log-normal distribution, with probability density function given in Appendix A, the model can be written as

$$Y_i|\theta_i \overset{ind}{\sim} \text{Poisson}(\theta_i)$$
$$\theta_i \overset{iid}{\sim} \mathrm{LN}(\mu, \sigma^2),$$

where $i = 1, \ldots, m$. Under this prior structure, the properties of conjugate distributions no longer apply, and therefore we have to tackle the problem of intractable mathematical integrations, in order to perform a Bayesian analysis. However, employing a log-normal prior distribution offers the advantage of more flexible modelling, especially when little prior information for the parameters of the first stage prior distribution is available. In this case, under the hierarchical Bayesian framework discussed later, a vague prior distribution on the first stage prior parameters, that is a distribution which does not favour any values of the parameters, would be a reasonable assumption. The gamma prior structure in (1.3) would generally require proper distributions on a and b (e.g. see Leonard and Novick, 1986, Christiansen and Morris, 1997, Daniels, 1999). On the other hand, the log-normal prior specification leads to proper posterior distributions, even under improper flat priors for the first stage parameters.

The assumption of a log-normal prior distribution for the Poisson means θ_i, $i = 1, \ldots, m$, can also be justified by the need to allow for more variation and possible outliers in the parameters θ_i, $i = 1, \ldots, m$, as illustrated in earlier work by Gaver and O'Muircheartaigh (1987), Carlin and Gelfand (1991) and Tierney (1994). Moreover, by employing the log-normal prior setting, one can assume a correlation structure in the Poisson means, which can be accommodated through the covariance matrix of the multivariate normal distribution of the reparametrised parameter vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_m)^{\mathrm{T}}$, where $\gamma_i = \log(\theta_i)$, $i = 1, \ldots, m$. An application of this use of the Poisson/log-normal model can be found in Clayton and Kaldor (1987) and Breslow and Clayton (1993).

1.4 Empirical Bayes estimation

To obtain posterior inferences for the parameters of interest, we first need to evaluate the unknown parameters of the prior distribution, as demonstrated in the case of the Poisson/gamma conjugate model, where the parameters of the posterior distribution were given earlier as functions of the hyperparameters in the updated $\mathrm{Ga}(a+y_i, b+1)$ distribution. As discussed in Chapter 3, the posterior distribution of θ_i under the Poisson/log-normal structure involves the parameters μ and σ^2 of the log-normal prior density, and thus these must be evaluated in order to be able to proceed with the posterior estimation of θ_i, $i = 1, \ldots, m$.

One can estimate the unknown parameters of the prior distribution, often referred to as the hyperparameters, following an empirical Bayes (EB) methodology (e.g. Morris, 1983a). According to this approach, the parameters of the last prior stage may be estimated using the observed data. Often, the method of moments based on the marginal density of the data, or the maximisation of the marginal likelihood is employed to obtain estimates of the hyperparameters. Gaver and O'-Muircheartaigh (1987) describe an EB analysis for a Poisson/log-normal model. Based on the EB approach, we obtain an estimator of the posterior mean of θ_i, $i = 1, \ldots, m$, which is derived as a minimum average risk approximation to $\mathrm{E}(\theta_i|\mathbf{y})$, expressed in the form of a linear shrinkage rule. We also suggest an estimator which relies on an importance sampling technique. Importance sampling methods (e.g. Geweke, 1989), generate random variates from an approximate distribution which is easier to simulate from, in relation to the original distribution of interest. Then, they suitably correct the sampled output, so that the generated values come from the distribution under consideration. The use of the importance sampling estimator in the Poisson/log-normal EB analysis was motivated by the resemblance of the conditional posterior distribution of θ_i, $i = 1, \ldots, m$, given the hyperparameters μ and σ^2, to a suitably tuned gamma distribution.

1.5 Hierarchical Bayes analysis

The EB methodology neglects the uncertainty associated with the hyperparameters. Accounting for this uncertainty would be a natural result of a full Bayesian modelling approach, as in that case it can be expressed through the assumption of a so-called second stage hyperprior distribution on the parameters μ and σ^2 of the log-normal prior. The second stage hyperprior distribution will again depend on some unknown parameters, and thus further hyperprior distributions can be assumed in a similar manner, forming a hierarchy of prior stages. The parameters at the lower level of the hierarchical model will either be considered known, or a vague prior specification may be assumed. In the hierarchical Bayesian analysis of the Poisson/log-normal model we consider two cases, assuming a vague prior setting for the hyperparameter μ, and both vague and informative prior information for the hyperparameter σ^2 of the log-normal distribution. The assumption of vague prior distributions can reflect a realistic and objective description of the problem under consideration when little prior knowledge is available. For the variance component σ^2, vague prior information is represented through a flat uniform distribution $U(0, \infty)$. This improper prior provides a proper posterior distribution for σ^2, unlike the uniform prior over the logarithm of σ^2, that

is $\log(\sigma^2) \sim U(-\infty, \infty)$ or equivalently $\pi(\sigma^2) \propto 1/\sigma^2$, which would lead to a nonintegrable posterior density (e.g. see Berger, 1985). A proper scaled inverse chi-square prior distribution could be another possibility, but the difficulty in the choice of its parameters in order to provide vague information makes the adoption of this prior more useful in the case when a more informative distribution is desired.

Under a hierarchical Bayes analysis of the Poisson/log-normal model, the posterior density of the Poisson means θ_i, $i = 1, \ldots, m$, is not given in a known form and does not permit further analytical integrations for the estimation of the characteristics of the posterior distribution. To overcome the problem of mathematical intractability we present some analytical approximations for the estimation of the posterior mean.

Numerical integration techniques, such as Gaussian quadrature, have also been used in the past to deal with such situations. However, this approach will still be inadequate in high dimensionality problems. Alternatively, inferences regarding the posterior distribution can also be drawn with the use of simulation techniques. The basic idea behind simulation-based inference methods, is to obtain a sufficient number of values generated from the distribution under consideration, and then use these values appropriately to estimate characteristics of interest. For example one may obtain the relevant moments and percentiles by suitable averaging and ordering of the simulation output, or the density function itself through a density estimation method. However, the nonstandard form of the posterior density of the parameters θ_i, $i = 1, \ldots, m$, in the hierarchical Poisson/log-normal model, does not allow direct simulation. In such cases, one can employ various rejection or importance sampling techniques to simulate from the distribution of interest. We investigate the possibility of adopting an importance sampling method, which similarly to the EB case, is based on sampling from a gamma distribution and attaching a suitable weight to each simulated value. The gamma density again serves as an approximation to the posterior distribution of $\gamma_i = \log(\theta_i)$, $i = 1, \ldots, m$.

Modern computing facilities offer an alternative tool for tackling intractable integrations involved in Bayesian analysis. The basic idea is to construct a Markov chain which has the posterior distribution of interest as its stationary distribution, simulate the chain until it converges to that stationary distribution, and then use an appropriate sample for Monte Carlo integration. Hence, in essence we iteratively simulate from distributions that eventually converge to the distribution under estimation. These so-called Markov chain Monte Carlo (MCMC) methods have become increasingly popular in recent years, especially

after Gelfand and Smith (1990) and Gelfand *et al.* (1990) illustrated their potential and general applicability in Bayesian inference. For the analysis of the hierarchical Poisson/log-normal model we consider the Gibbs sampler, which is an MCMC method that iteratively simulates from the full conditional posterior distributions of each model parameter, given all the remaining parameters in the model. Nevertheless, the full conditional posterior densities of the Poisson means $\theta_i,\ i = 1, \ldots, m$, in the considered model do not correspond to any known distribution, and therefore they cannot be sampled directly. We tackle this problem by developing a log-normal/gamma mixture approximation to $p(\theta_i|\mu, \sigma^2, \mathbf{y})$ in such a way that it has the same first three moments as the exact distribution, and is based on a discretisation of the prior log-normal distribution and an entropy distance minimisation method. We also suggest a hybrid MCMC algorithm, which corrects the mixture approximation involved in the Gibbs sampling scheme by incorporating a Metropolis-Hastings acceptance step in each Gibbs sampling iteration.

1.6 Overview of thesis

In Chapter 2 we review methods that have been suggested in the literature for the estimation of several Poisson means. We present various linear shrinkage estimators, which are developed under a frequentist philosophy and are constructed with the objective of dominating the unbiased estimator in terms of the risk under a specific loss function. Then, the Bayesian approach is considered, and empirical Bayes methods derived as linear shrinkage rules are investigated.

Chapter 3 begins with the introduction of the Bayesian Poisson/log-normal formulation, and then the posterior inference under this structure is discussed. We derive two approximations to the posterior mean of the parameters of interest $\theta_i,\ i = 1, \ldots, m$. The first is a linear estimator which minimises the average risk with respect to a summed quadratic loss function. The second is a nonlinear estimator based on Monte Carlo integration through a suitable importance sampling technique. The implementation of both methods requires knowledge of the parameters of the prior $\mathrm{LN}(\mu, \sigma^2)$ distribution, and therefore the EB estimation methodology is considered. We illustrate the methods using real data examples, and the shrinking behaviour of the EB estimators is explored by means of simulated data. We also examine the frequency properties of the EB methods, assessing their average risk through repeated simulations, and we compare them with various methods proposed in the literature.

In Chapter 4 we present a hierarchical Poisson/log-normal model, by intro-

ducing two different structures comprising a further prior stage for the log-normal parameters μ and σ^2. We first assume that both the hyperparameters are distributed according to independent flat uniform distributions, and then we consider that while μ is again uniformly distributed over its range, the variance parameter σ^2 independently follows an informative scaled inverse chi-square hyperprior distribution with suitably chosen parameters. We then derive some analytical approximations for the posterior mean of the Poisson means θ_i, $i = 1, \ldots, m$, under both the conjugate Poisson/gamma and the considered Poisson/log-normal formulations.

Chapter 5 contains simulation-based methods for the analysis of the hierarchical Poisson/log-normal model. We first attempt to estimate the Poisson parameters, under both hyperprior settings for μ and σ^2, employing a noniterative importance sampling Monte Carlo integration technique, and the reliability of the method is assessed. An MCMC approach is then adopted that employs the Gibbs sampling algorithm. We derive a log-normal/gamma approximation to the full conditional posterior distribution of the Poisson means θ_i, $i = 1, \ldots, m$, that allows direct simulation for the implementation of the Gibbs sampler. We also suggest drawing exact posterior inferences by using a hybrid MCMC strategy which combines the Gibbs sampler with a Metropolis-Hastings algorithm, and is based on the same mixture approximation as before. Applications of the proposed methods are given, and the shrinking behaviour of the hierarchical Bayes estimator is empirically assessed.

In Chapter 6 we report the results from frequency simulations, in order to investigate the average risk of the estimators resulting from the full hierarchical Bayes analysis. Finally, the use of the developed methods in possible extensions to the considered models is explored and illustrated by means of relevant applications.

Chapter 2

Shrinkage estimators of several Poisson means

2.1 Introduction

We consider the problem of the multiparameter Poisson estimation, assuming that given the parameters $\theta_1, \theta_2, \ldots, \theta_m$, the counts $Y_1, Y_2, \ldots, Y_m$, are independent Poisson variables with respective means $\theta_1, \theta_2, \ldots, \theta_m$, according to model (1.1). The aim is to draw inferences for the means of the m conditionally independent Poisson distributions, θ_i, $i = 1, \ldots, m$, simultaneously.

The problem of the simultaneous multiparameter estimation was initially addressed in the case of normal means. Stein (1956) showed that the usual estimator for the mean vector parameter, that is the MLE, is inadmissible with respect to the squared error loss function when the number of means exceeds two. James and Stein (1961) provided an estimator which dominates the MLE in terms of mean squared error (MSE), and Stein (1973) suggested an integration by parts technique for obtaining estimators with improved MSE properties when compared to the usual estimator. Based on this result, various methods have been proposed for tackling the problem of the multiparameter estimation in the normal case, as well as for the continuous exponential family in general (e.g. Hudson, 1978).

Returning to the Poisson case in model (1.1), the usual estimator for the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\mathrm{T}}$, is given by $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)^{\mathrm{T}}$, that being the MLE and the UMVUE, as mentioned in Chapter 1. However, if our objective is to obtain estimators with good MSE properties, we may be prepared to allow for some bias in order to decrease the variance of the estimator and obtain smaller MSE.

Definition 2.1. *Let $\boldsymbol{\theta}^* = (\theta_1, \theta_2, \ldots, \theta_m)^T$ be an estimator of the parameter vector $\boldsymbol{\theta}$, and $L(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ denote a loss function, that is a 'cost' involved when $\boldsymbol{\theta}$ is imprecisely estimated by $\boldsymbol{\theta}^*$. Then, the risk $R(\boldsymbol{\theta}^*)$ associated with the loss function*

$L(\boldsymbol{\theta}^*, \boldsymbol{\theta})$, is defined as the expected value of $L(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ with respect to the conditional distribution of Y given θ, i.e.

$$R(\boldsymbol{\theta}^*) = E_{Y|\theta}\{L(\boldsymbol{\theta}^*, \boldsymbol{\theta})\}.$$

Now, if we let

$$\text{SEL}_k = \sum_{i=1}^{m} \frac{(\theta_i^* - \theta_i)^2}{\theta_i^k}$$

denote a weighted summed squared error loss function for a given nonnegative integer value of k, when the parameter θ_i is estimated by θ_i^*, then it is obvious that $R(\boldsymbol{\theta}^*)$ gives the MSE of an estimator when $L(\boldsymbol{\theta}^*, \boldsymbol{\theta}) \equiv \text{SEL}_0$.

Definition 2.2. *An estimator $\boldsymbol{\theta}^*$ of $\boldsymbol{\theta}$ is inadmissible if there exists an estimator $\boldsymbol{\theta}'$ such that*

$$R(\boldsymbol{\theta}') \leq R(\boldsymbol{\theta}^*)$$

for all the parameter vectors $\boldsymbol{\theta}$, with strict inequality for at least one value of $\boldsymbol{\theta}$.

Peng (1975) first proved that the MLE is inadmissible with respect to the SEL_0 loss function for the simultaneous estimation of the means of several independent Poisson distributions, given that the number of the means is at least equal to three. Hudson (1978) provides improved estimators under the same loss function for the multiparameter estimation for distributions in the discrete exponential family. Other authors (e.g. see Ghosh, Hwang and Tsui, 1983) consider different loss functions and construct various estimators which universally dominate the MLE in terms of the risk associated with the loss function under consideration. Also, Efron and Morris (1973), Leonard (1976), Albert (1981) and others, have proposed estimators for which the improvement in comparison to the MLE over the entire parameter space cannot be proven theoretically. These authors argue that such an improvement is of no practical use, since the statistician will normally be interested in a certain region within which the parameter $\boldsymbol{\theta}$ is likely to lie.

The estimators considered in this chapter can take the general form

$$\boldsymbol{\delta}(\mathbf{Y}) = \mathbf{Y} + \mathbf{g}(\mathbf{Y}) \tag{2.1}$$

with $\boldsymbol{\delta}(\mathbf{Y}) = \{\delta_1(\mathbf{Y}), \delta_2(\mathbf{Y}), \ldots, \delta_m(\mathbf{Y})\}^{\mathrm{T}}$, and $\mathbf{g}(\mathbf{Y}) = \{g_1(\mathbf{Y}), g_2(\mathbf{Y}), \ldots, g_m(\mathbf{Y})\}^{\mathrm{T}}$ being a function to be specified later. They are known as shrinkage estimators since, depending on the nature of the function $\mathbf{g}(\mathbf{Y})$, they shrink the usual estimate $\mathbf{Y}$ towards a fixed or data-determined point. Various suggestions for the point of shrinkage include zero, the smallest observation of the data, the sample

mean, the geometric mean, the median, an arbitrary prior guess etc. Clearly, when $\mathbf{g}(\mathbf{Y}) \neq 0$, any estimator given by (2.1) is no longer an unbiased estimator. By smoothing the usual estimate $\mathbf{Y}$ towards a fixed or data-based point we introduce some bias in the estimates. However, at the same time we expect the variance to be reduced, resulting in estimates that will have better risk properties. In essence, with shrinkage estimators the inference regarding any single component of the parameter vector $\boldsymbol{\theta}$ is based on adjusting the corresponding component of the unbiased estimator $\mathbf{Y}$, using information contained in the entire vector $\boldsymbol{\theta}$ or $\mathbf{Y}$. Also, by summing over the $\boldsymbol{\theta}$ components in the loss function, e.g. by employing a loss structure of the form $\mathrm{SEL}_k = \sum_{i=1}^m \{\delta_i(\mathbf{Y}) - \theta_i\}^2/\theta_i^k$, for a nonnegative integer k, we exploit the strength in all the elements of $\boldsymbol{\theta}$ as far as the risk evaluation is concerned. This offers an intuitive explanation to the fact that in multiparameter estimation we can improve over the risk properties of the MLE, although the latter is admissible when only one parameter is to be estimated (e.g. see Hodges and Lehmann, 1951).

2.1.1 Inadmissibility of the MLE

The possibility of improving upon the UMVUE in terms of risk in the problem of the simultaneous estimation of several Poisson means has been investigated by many authors. They consider various loss functions and show that estimators with universally smaller risk than that of the MLE can be obtained. Peng (1975) and Hudson (1978) consider the squared error loss function $\mathrm{SEL}_0 = \sum_{i=1}^m \{\delta_i(\mathbf{Y}) - \theta_i\}^2$, Clevenson and Zidek (1975) obtain improved estimators under the normalised loss $\mathrm{SEL}_1 = \sum_{i=1}^m \{\delta_i(\mathbf{Y}) - \theta_i\}^2/\theta_i$, while Tsui and Press (1982) and Ghosh, Hwang and Tsui (1983) employ more general loss functions of the form $\mathrm{SEL}_k = \sum_{i=1}^m \{\delta_i(\mathbf{Y}) - \theta_i\}^2/\theta_i^k$, $k = 0, 1, 2$, and prove the inadmissibility of the MLE. The derived estimators take a common similar form, but differ mainly according to the point towards which the smoothing is directed. Shrinkage estimators are expected to exhibit low risk when they shift the usual estimate towards a region where the true parameter is likely to lie. We can distinguish them to fixed-point shrinkage estimators, which shrink the MLE towards some predetermined fixed point, and the so-called adaptive shrinkage estimators, for which the direction of shrinking depends on the observed data.

We will now demonstrate the inadmissibility of the MLE in the multiparameter Poisson estimation problem, by presenting a certain class of shrinkage rules that universally dominates this estimator under the summed squared error loss function SEL_0. The arguments here correspond to a method resulting in fixed-point estimators.

We let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)^{\mathrm{T}}$ be the vector of m conditionally independent Poisson random variables, given their respective means $\theta_1, \theta_2, \ldots, \theta_m$. We consider estimators of the form given in (2.1), that is $\boldsymbol{\delta}(\mathbf{Y}) = \mathbf{Y} + \mathbf{g}(\mathbf{Y})$, where

$$\mathbf{g}(\mathbf{Y}) = \{g_1(\mathbf{Y}), g_2(\mathbf{Y}), \ldots, g_m(\mathbf{Y})\}^{\mathrm{T}}, \tag{2.2}$$

with $g_i(\mathbf{Y})$, $i = 1, \ldots, m$, being real-valued functions such that

$$\begin{aligned} g_i(y) &= 0, \qquad \text{if} \qquad y \leq 0 \\ \mathrm{E}|g_i(\mathbf{Y})| &< \infty. \end{aligned} \tag{2.3}$$

Then, under the squared error loss function $\mathrm{SEL}_0 = \sum_{i=1}^{m}\{\delta_i(\mathbf{Y}) - \theta_i\}^2$, the difference in risk between the MLE and $\boldsymbol{\delta}(\mathbf{Y})$ is given by

$$\begin{aligned} \Delta R &= R(\mathbf{Y}) - R\{\boldsymbol{\delta}(\mathbf{Y})\} \\ &= \mathrm{E}_{Y|\theta}\left\{\sum_{i=1}^{m}(Y_i - \theta_i)^2\right\} - \mathrm{E}_{Y|\theta}\left[\sum_{i=1}^{m}\{Y_i + g_i(\mathbf{Y}) - \theta_i\}^2\right] \\ &= \mathrm{E}_{Y|\theta}\{\psi(\mathbf{Y})\}, \end{aligned} \tag{2.4}$$

where

$$\psi(\mathbf{Y}) = \sum_{i=1}^{m}(Y_i - \theta_i)^2 - \sum_{i=1}^{m}\{Y_i + g_i(\mathbf{Y}) - \theta_i\}^2 .$$

Expanding the squares in the previous expression, we can write

$$\psi(\mathbf{Y}) = -2\sum_{i=1}^{m}\{Y_i g_i(\mathbf{Y}) - g_i(\mathbf{Y})\theta_i\} - \sum_{i=1}^{m}\{g_i(\mathbf{Y})\}^2 . \tag{2.5}$$

The following lemma will allow us to eliminate the parameter θ_i from (2.5) when we take the expectation of $\psi(\mathbf{Y})$ with respect to the conditional distribution of Y given θ. It is the discrete equivalent of Stein's integration by parts result (Stein, 1973).

Lemma 2.1. *Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)^T$, be a vector consisting of independent Poisson random variables with means $\theta_1, \theta_2, \ldots, \theta_m$, and $g(\cdot)$ be a real-valued function such that $g(y) = 0$ if $y \leq 0$ and $E|g(y)| < \infty$. We also let $\mathbf{e}_i$ denote the m-dimensional vector having the i-th coordinate equal to one and all the others equal to zero. Then*

$$E\{Y_i\, g(\mathbf{Y} - \mathbf{e}_i)\} = E\{g(\mathbf{Y})\, \theta_i\} . \tag{2.6}$$

Proof. We can write

$$\mathrm{E}\{Y_i\ g(\mathbf{Y}-\mathbf{e}_i)\} = \begin{pmatrix} \mathrm{E}\{Y_i\ g(Y_1)\} \\ \vdots \\ \mathrm{E}\{Y_i\ g(Y_i-1)\} \\ \vdots \\ \mathrm{E}\{Y_i\ g(Y_m)\} \end{pmatrix}, \quad i=1,\ldots,m. \tag{2.7}$$

We now notice that

$$\mathrm{E}\{Y_i\ g(Y_j)\} = \mathrm{E}(Y_i)\ \mathrm{E}\{g(Y_j)\} = \theta_i\ \mathrm{E}\{g(Y_j)\}, \qquad \text{for} \qquad i \neq j. \tag{2.8}$$

When $i = j$ we have

$$\begin{aligned} \mathrm{E}\{Y_i\ g(Y_i-1)\} &= \sum_{y_i=0}^{\infty} y_i\ g(y_i-1)\ \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!} \\ &= 0 + \sum_{y_i=1}^{\infty} g(y_i-1)\ \frac{e^{-\theta_i}\theta_i^{y_i}}{(y_i-1)!} \\ &= \sum_{y_i=0}^{\infty} \theta_i\ g(y_i)\ \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!} \\ &= \mathrm{E}\{\theta_i\ g(Y_i)\} = \theta_i\ \mathrm{E}g(Y_i). \end{aligned} \tag{2.9}$$

Using (2.8) and (2.9), it is obvious that (2.7) can be written as

$$\mathrm{E}\{Y_i\ g(\mathbf{Y}-\mathbf{e}_i)\} = \begin{pmatrix} \theta_i\ \mathrm{E}g(Y_1) \\ \vdots \\ \theta_i\ \mathrm{E}g(Y_i) \\ \vdots \\ \theta_i\ \mathrm{E}g(Y_m) \end{pmatrix} = \mathrm{E}\{g(\mathbf{Y})\ \theta_i\}\,.$$

□

Returning to the difference in risk between the MLE and $\boldsymbol{\delta}(\mathbf{Y})$, we notice that using the result (2.6), Equations (2.4) and(2.5) give

$$\Delta R = \mathrm{E}_{Y|\theta}\{\psi(\mathbf{Y})\} = -2\ \mathrm{E}_{Y|\theta}\{\mathfrak{D}(\mathbf{Y})\}\,, \tag{2.10}$$

with $\mathfrak{D}(\mathbf{Y})$ given as

$$\begin{aligned} \mathfrak{D}(\mathbf{Y}) &= \sum_{i=1}^{m} Y_i\ \{g_i(\mathbf{Y}) - g_i(\mathbf{Y}-\mathbf{e}_i)\} + \frac{1}{2}\ \sum_{i=1}^{m}\{g_i(\mathbf{Y})\}^2 \\ &= \sum_{i=1}^{m}\{Y_i\ \Delta_i g_i(\mathbf{Y})\} + \frac{1}{2}\ \sum_{i=1}^{m}\{g_i(\mathbf{Y})\}^2, \end{aligned} \tag{2.11}$$

where for a function $f(\cdot)$ and a vector $\mathbf{X}$ we let $\Delta_i f(\mathbf{X})$ denote the difference $f(\mathbf{X}) - f(\mathbf{X} - \mathbf{e}_i)$.

Clearly, for the shrinkage estimator $\boldsymbol{\delta}(\mathbf{Y})$ to dominate the MLE we need the difference in risk given in (2.10) to be nonnegative for all $\mathbf{Y}$ and positive for some $\mathbf{Y}$. This follows if

$$\mathfrak{D}(\mathbf{Y}) \leq 0 \qquad \text{for all} \qquad \mathbf{Y}, \tag{2.12}$$

with strict inequality for some $\mathbf{Y}$. This is true for estimators shrinking $\mathbf{Y}$ towards a vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_m)^{\mathrm{T}}$, for which the solution to the difference inequality (2.12) is of the form

$$g_i(\mathbf{Y}) = -\frac{C(\mathbf{Y})\ H_i(Y_i)}{D(\mathbf{Y})}, \tag{2.13}$$

where $C(\mathbf{Y}) \geq 0$, $D(\mathbf{Y}) = \sum_{j=1}^{m} d_j(Y_j)$ with $d_j(Y_j)$ being a nonnegative function of Y_j, and H_i has the form

$$H_i(Y_i) = h_i(Y_i) - h_i(\lambda_i)$$

where

$$h_i(Y_i) = \sum_{j=1}^{Y_i} \frac{1}{j}.$$

The following theorem is similar to one given in Hwang (1982) and provides a nonpositive upper bound for $\mathfrak{D}(\mathbf{Y})$. The proof is given in Appendix B.

Theorem 2.1. *Under the assumptions given in Appendix B, the vector* $\mathbf{g}(\mathbf{Y}) = \{g_1(\mathbf{Y}), g_2(\mathbf{Y}), \ldots, g_m(\mathbf{Y})\}^T$ *with* $g_i(\mathbf{Y})$ *given by (2.13), satisfies (2.12), where for all* $\mathbf{Y}$, $C(\mathbf{Y})$ *is such that*

$$H_i(Y_i - 1)\ \Delta_i C(\mathbf{Y}) \geq 0, \tag{2.14}$$

$$\textit{and} \qquad 0 \leq C(\mathbf{Y}) \leq K^{-1}\{N(Y) - \beta\}_+, \tag{2.15}$$

where $N(Y)$ *is the number of* Y_j, $j = 1, \ldots, m$, *that exceed* λ_i *and* $\{\cdot\}_+$ *denotes the positive part function. Furthermore, the difference* $\mathfrak{D}(\mathbf{Y})$ *satisfies the inequality*

$$\mathfrak{D}(\mathbf{Y}) \leq -\frac{C(\mathbf{Y})\ \{N(Y) - \beta - K\ C(\mathbf{Y})\}_+}{D(\mathbf{Y})}, \tag{2.16}$$

with the strict inequality holding for the vectors $\mathbf{Y}$ *satisfying*

$$C(\mathbf{Y})\ H_i(Y_i - 1)\ \Delta_i d_i(Y_i) > 0 \qquad \textit{for at least two } i. \tag{2.17}$$

Theorem 2.1 shows that a certain class of estimators that dominate the MLE in terms of risk under the squared error loss function $\text{SEL}_0 = \sum_{i=1}^{m}\{\delta_i(\mathbf{Y}) - \theta_i\}^2$ can be found. It can be extended to the case where the more general loss function $\text{SEL}_k = \sum_{i=1}^{m}\{\delta_i(\mathbf{Y})-\theta_i\}^2/\theta_i^k$, for integer k, is considered (Hwang, 1982). The exact form of the estimators is determined by the form of the functions $C(\mathbf{Y})$, $H_i(Y_i)$ and $D(\mathbf{Y})$, subject to the conditions (2.14) and (2.15) and the assumptions stated in Appendix B. We notice that the improvement over the MLE has an upper bound, given in (2.16), which depends on the functions $C(\mathbf{Y})$ and $D(\mathbf{Y})$ that form the estimator. For instance, a small $D(\mathbf{Y})$ would be desirable to obtain greater improvement in risk compared to the usual estimator. Universal dominance of shrinkage estimators over the MLE under other loss functions can also be shown. In the subsequent sections we present a brief review of some methods that have appeared in the literature for the simultaneous estimation of several independent Poisson means.

2.2 Linear shrinkage estimators

We will now describe some estimators having the form (2.1) with the components of the vector $\mathbf{g}(\mathbf{Y})$ being of the type given in (2.13). We first introduce some general notation that will be used to present the various methods.

Notation:

(i) $\{\cdot\}_+ = \max\{0, \cdot\}$;

(ii) $N(j)$: number of Y_i, $i = 1, \ldots, m$, such that $Y_i > j$;

(iii) $h(Y_i) = \begin{cases} \sum_{j=1}^{Y_i} \frac{1}{j}, & \text{when} \quad Y_i \geq 1; \\ 0, & \text{otherwise;} \end{cases}$

(iv) $\bar{h} = \sum_{i=i}^{m} \frac{h(Y_i)}{m}$;

(v) $Y_{(i)}$: ith ordered observation;

(vi) Y_{med} : median of data.

The described estimators adjust the MLE towards a chosen fixed or data-based point, that depending on the nature of the function $H_i(Y_i)$. They are also constructed in such a way, that they dominate the usual estimator under a given loss function.

2.2.1 Peng estimator

Peng (1975) showed that there exist estimators that dominate the MLE under $\text{SEL}_0 = \sum_{i=1}^{m}\{\delta_i(\mathbf{Y}) - \theta_i\}^2$. He obtained the following estimator

$$\delta_i^P(\mathbf{Y}) = Y_i - \{N(0) - 2\}_+ \frac{h(Y_i)}{\sum_{i=1}^{m} h^2(Y_i)} \tag{2.18}$$

which improves the risk of the unbiased estimator under SEL_0 when $m \geq 3$. His method shrinks towards zero and therefore should be expected to provide considerably lower risk than the MLE when the Poisson means are close to the origin. We notice that the transformation $h(Y) = \sum_{j=1}^{Y} \frac{1}{j}$ corresponds to the square root transformation for small Y and to the logarithm transformation for large Y. These two transformations are widely used to approximate Poisson observations with normally distributed data. Therefore, considering the James and Stein (1961) estimator for normal data $\mathbf{X} = (X_1, X_2, \ldots, X_m)^{\mathrm{T}}$, given as

$$\delta_i(\mathbf{X}) = X_i - \frac{(m-2)\, X_i}{\sum_{i=i}^{m} X_i^2}$$

we can clearly see the intuition behind Peng's estimator. In fact, the same intuition and similarity to various Stein-type methods, is present in most of the estimators described in this chapter.

Tsui (1981) and Hudson and Tsui (1981) suggested more general estimators that smooth towards a nonnegative prior guess λ, which reduce to Peng's method when $\lambda = 0$.

2.2.2 Tsui estimator

The methods given by Tsui (1981) and Hudson and Tsui (1981), mentioned in the preceding subsection, require that some prior guess of the true Poisson means can be obtained. However, for the case that such prior knowledge is not available, Tsui (1981) derived a shrinkage rule similar to that of Peng, which adjusts the data towards the direction of the minimum observation. The estimator is given by

$$\delta_i^T(\mathbf{Y}) = Y_i - \{N(Y_{(1)}) - 2\}_+ \frac{h_{Y_{(1)}}(Y_i)}{\sum_{i=1}^{m} h_{Y_{(1)}}^2(Y_i)} \tag{2.19}$$

where

$$h_{Y_{(1)}}(Y_i) = \begin{cases} 1 + \sum_{j=2}^{Y_i - Y_{(1)}} \frac{1}{Y_{(1)}+j}, & \text{if} \quad Y_i \geq Y_{(1)} + 1 \\ 0, & \text{otherwise.} \end{cases}$$

This estimator dominates the MLE under SEL_0 when the number of Poisson means exceeds three.

2.2.3 Ghosh, Hwang and Tsui estimators

Ghosh, Hwang and Tsui (1983) obtained several estimators which are constructed under specific forms of the more general weighted loss function $\text{SEL}_k = \sum_{i=1}^{m}\{\delta_i(\mathbf{Y}) - \theta_i\}^2/\theta_i^k$, with k being a nonnegative integer, and shrink to possibly different directions. Under the squared error loss SEL_0 they suggest using

$$\delta_i^{GHT1}(\mathbf{Y}) = Y_i - \{N(Y_{(1)}) - 2\}_+ \frac{h(Y_i) - h(Y_{(1)})}{\sum_{i=1}^{m}\{h(Y_i) - h(Y_{(1)})\}\{h(Y_i+1) - h(Y_{(1)})\}} \tag{2.20}$$

which pools the data towards the minimum observation and dominates the MLE for $m \geq 4$. This method is similar to Tsui's estimator in (2.19).

When the normalised squared error loss function $\text{SEL}_1 = \sum_{i=1}^{m}\{\delta_i(\mathbf{Y}) - \theta_i\}^2/\theta_i$ is considered, Ghosh, Hwang and Tsui (1983) propose the estimator

$$\delta_i^{GHT2}(\mathbf{Y}) = Y_i - \{N(Y_{(1)}) - 1\}_+ \frac{Y_i - Y_{(1)} - 1}{\sum_{i=1}^{m}\{Y_i - Y_{(1)}\} - 1} \tag{2.21}$$

which dominates the MLE when three or more Poisson means are to be estimated. The shrinkage of rule (2.21) is again towards the minimum observation in the data. However, it can be easily modified to smooth towards any ordered observation according to a more general result given by the same authors.

All the estimators presented until now adjust the MLE by shrinking towards zero or the first ordered observation. Despite the fact that they all dominate the usual estimator with respect to some specific loss function, the relative savings in risk are not expected to be considerable when the true means are large. To account for such situations, Ghosh, Hwang and Tsui (1983) consider an estimator that shrinks the MLE towards the median of the data, given by

$$\delta_i^{GHT3}(\mathbf{Y}) = Y_i - \{N(Y_{med}) - 2\}_+ \frac{h(Y_i) - h(Y_{med})}{\sum_{i=1}^{m} d_i(Y_i)} \tag{2.22}$$

where

$$d_i(Y_i) = \begin{cases} \{h(Y_i) - h(Y_{med})\}^2 + \frac{1}{2}\{3h(Y_{med}) - 2\}_+, & \text{if } Y_i < Y_{med} \\ \{h(Y_i) - h(Y_{med})\}\{h(Y_i+1) - h(Y_{med})\}, & \text{if } Y_i \geq Y_{med}. \end{cases}$$

This rule has universally smaller risk than the usual UMVUE estimator under SEL_0 when $m \geq 6$.

The same authors propose another method that smoothes towards some central point of the data. We let $N_h(\bar{h})$ denote the number of i for which $h(Y_i) > \bar{h}$. Then, the estimator

$$\delta_i^{GHT4}(\mathbf{Y}) = Y_i - \{N_h(\bar{h}) - 2\}_+ \frac{h(Y_i) - \bar{h}}{\sum_{i=1}^{m}\{h(Y_i) - \bar{h}\}^2} \tag{2.23}$$

shrinks the MLE towards $\bar{h} = \sum_{j=1}^{m} h(Y_j)/m$. But, for large Y_j, $h(Y_j) \doteq \log(Y_j)$, and therefore $\bar{h}$ approximates the geometric mean of the data. Thus, estimator (2.23) tends towards the geometric mean when the observations are large. However, the dominance of this method over the MLE has only been shown empirically.

2.2.4 Hudson estimator

Hudson (1985) presents a method that is similar to estimator (2.23) in shrinking towards a value close to the geometric mean of the data. He shows that the estimator given by

$$\delta_i^H(\mathbf{Y}) = \begin{cases} Y_i - \{N(0) - 3\}_+ \frac{h(Y_i) - \bar{h}}{\sum_{j=1}^{m} \{h(Y_j) - \bar{h}\}^2}, & \text{if } Y_i + 0.56 > \frac{\{N(0)-3\}_+}{\sum_{i=j}^{m} \{h(Y_j) - \bar{h}\}^2} \\ 0.56\ \{\exp(\bar{h}) - 1\}, & \text{otherwise} \end{cases} \tag{2.24}$$

approximately dominates the usual estimator $\mathbf{Y}$ for large m under the squared error loss SEL_0. We may also view the estimator (2.24) as a rule smoothing towards the fitted values of a reduced log-linear model. The transformation $h(Y_i) = \sum_{j=1}^{Y_i} \frac{1}{j}$, if $Y_i \geq 1$, and $h(0) = 0$, can be approximated by the log-like transformation $\log\left(\frac{Y_i + 0.56}{0.56}\right)$, and therefore under this transformation of the data we have a form of a log-linear model. In that case, estimator (2.24) smoothes the transformed data values $h(Y_i)$ towards the fitted values $\bar{h}$ of an exchangeable model of a log-linear form. The condition in (2.24) ensures that the shrinkage does not exceed the fitted value.

2.2.5 Clevenson and Zidek estimator

Clevenson and Zidek (1975) developed estimators that dominate the MLE under the normalised squared error loss function $\text{SEL}_1 = \sum_{i=1}^{m} \{\delta_i(\mathbf{Y}) - \theta_i\}^2/\theta_i$. They show that the estimator

$$\delta_i^{CZ}(\mathbf{Y}) = Y_i - \frac{(\beta + m - 1)\ Y_i}{\sum_{i=1}^{m} Y_i + \beta + m - 1} \tag{2.25}$$

with $0 \leq \beta \leq m - 1$, outperforms $\mathbf{Y}$ in terms of risk under SEL_1 when $m \geq 2$. In the special case where $\beta = 1$, (2.25) is the same as one of the estimators proposed by Tsui and Press (1982). This method shrinks towards the origin and should therefore be expected to perform well when the Poisson means θ_i are small. Moreover, we notice that in the Clevenson and Zidek estimator, with $\beta = 1$, the function $D(\mathbf{Y})$ in (2.16) is given by $D(\mathbf{Y}) = \sum_{i=1}^{m} Y_i + m$, whereas for the estimator $\delta_i^{GHT2}(\mathbf{Y})$ in (2.21), which is also developed under SEL_1, $D(\mathbf{Y})$ is

equal to $\sum_{i=1}^{m}\{Y_i - Y_{(1)}\} - 1$. This implies that the latter method will produce greater relative savings in risk when the Poisson means are large but close to each other.

2.2.6 Tsui and Press estimator

Until now we have described improved estimators under the loss functions SEL_0 and SEL_1. The former is widely used in statistical decision making problems, whereas the latter may be useful when poor estimation for very small θ_i is highly undesirable. Another intuitive reason for choosing the normalised squared error loss function SEL_1, is that in the case of the simultaneous estimation of m independent normal means, using the random variables $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, m$, the loss function $\sum_{i=1}^{m}(\mu_i - \hat{\mu}_i)^2/\sigma_i^2$, would seem a natural choice. However, if the inference is focused on the variance parameters σ_i^2, a reasonable loss function could be

$$\sum_{i=1}^{m}\left\{1 - \frac{\hat{\sigma}_i^2}{\sigma_i^2}\right\}^2 = \sum_{i=1}^{m}\frac{(\sigma_i^2 - \hat{\sigma}_i^2)^2}{\sigma_i^4}.$$

Then, for the Poisson case one could consider the loss function $\text{SEL}_2 = \sum_{i=1}^{m}\{\delta_i(\mathbf{Y}) - \theta_i\}^2/\theta_i^2$ for the parameter θ_i, which also gives the variance of the Poisson distribution. Tsui and Press (1982) show that under this weighted squared error loss function, the estimator

$$\delta_i^{TP}(\mathbf{Y}) = Y_i - 2(m-1)\,\frac{Y_i\,(Y_i - 1)}{\sum_{\substack{j=1 \\ j\neq i}}^{m}(Y_j + 2)\,(Y_j + 1) + Y_i\,(Y_i - 1)} \tag{2.26}$$

dominates the MLE when $m \geq 2$. Their method adjusts the usual estimator towards zero.

2.3 Bayesian approach to shrinkage estimation

In Section 2.2 we did not assume any prior knowledge on the Poisson means θ_i. Let us now consider a general situation where given the parameters $\theta_1, \theta_2, \ldots, \theta_m$, the random variables $Y_1, Y_2, \ldots, Y_m$, independently follow a sampling distribution with mean θ_i and variance $V_i(\theta_i)$, for $i = 1, \ldots, m$. We further assume that the parameters θ_i have independent prior distributions with mean M_i and variance A_i, $i = 1, \ldots, m$. If we let the notation $Y \sim [\text{E}(Y),\ \text{var}(Y)]$ indicate that the random variable Y has mean $\text{E}(Y)$ and variance $\text{var}(Y)$, with no further assumptions for its distribution, we can write

$$Y_i|\theta_i \sim [\theta_i,\ V_i(\theta_i)] \tag{2.27}$$

$$\theta_i \sim [M_i,\ A_i], \tag{2.28}$$

for $i = 1, \ldots, m$. We also let G_i denote the prior expectation of the variance of Y_i given θ_i, that is

$$G_i = \mathrm{E}_\theta\{\mathrm{var}(Y_i|\theta_i)\} = \mathrm{E}_\theta\{V_i(\theta_i)\}.$$

Then, Ericson (1969) and Goldstein (1975) show that if the conditional posterior expectation of θ_i can be written in the linear form

$$\mathrm{E}(\theta_i|\mathbf{y}) = aY_i + b, \tag{2.29}$$

then

$$a = \frac{A_i}{A_i + G_i} \qquad \text{and} \qquad b = (1-a)M_i, \tag{2.30}$$

implying that the posterior mean is a shrinkage estimator for θ_i, adjusting the MLE towards the prior mean M_i according to a weight given by $(1-a)$, which is inversely proportional to the prior variance A_i. An exact linear form (2.29) can be assumed for the posterior mean, as long as the prior distribution of θ_i is the conjugate with respect to the conditional distribution of Y_i given θ_i.

2.3.1 Empirical Bayes

The quantities a and b in the linear estimator (2.29) are given as functions of the moments of the prior distribution of the Poisson means θ_i, $i = 1, \ldots, m$. In order to be able to use the estimator (2.29), the normally unknown moments M_i, A_i and G_i need to be evaluated. Several authors suggest to estimate them using the marginal distribution of the data, adopting an EB approach.

Efron and Morris type estimators

Under the general distributional assumptions (2.27) and (2.28), Efron and Morris (1973) suggest the linear shrinkage estimator

$$\delta_i(\mathbf{Y}) = (1 - C_{EM})\ Y_i + C_{EM}\ M_i \tag{2.31}$$

which smoothes the usual estimate Y_i of the parameter θ_i towards its prior mean M_i with a weight equal to

$$C_{EM} = \frac{G_i}{A_i + G_i}.$$

This means that, according to (2.29) and (2.30), the estimator proposed by Efron and Morris is the posterior mean of θ_i, whenever the later can take the linear form (2.29). Efron and Morris (1973) show that the rule given in (2.31) dominates the

MLE under the squared error loss SEL_0 asymptotically, that is when $m \to \infty$. However, it is expected to provide considerable improvement in risk when the parameter $\boldsymbol{\theta}$ lies in a region close to its prior mean. The authors suggest an EB solution when M_i and C_{EM} are unknown.

More specifically, Leonard (1976) assumes a Poisson(θ_i) distribution in (2.27) and tackles the problem of the simultaneous estimation of several Poisson means from a Bayesian perspective. He derives the estimator

$$\delta_i^L(\mathbf{Y}) = (1 - C_L)\ Y_i + C_L\ \overline{Y} \tag{2.32}$$

where the shrinkage proportion C_L is obtained in a way such that it minimises the risk of the estimator (2.32) under the squared error loss function SEL_0, and is estimated from the data as

$$C_L = \min\left\{\frac{\overline{Y}}{s^2},\ 1\right\}, \tag{2.33}$$

where s^2 denotes the sample variance, that is $s^2 = \sum_{i=1}^m (Y_i - \overline{Y})^2/(m-1)$. Clearly, this EB rule shrinks the usual estimate Y_i towards, but not beyond, the sample mean $\overline{Y}$ of the data. We notice here that the risk of the MLE under the squared error loss SEL_0 is given by

$$\begin{aligned} R(\mathbf{Y}) &= \text{E}_{Y|\theta}\left\{\sum_{i=1}^m (Y_i - \theta_i)^2\right\} \\ &= \sum_{i=1}^m \text{var}_{Y|\theta}(Y_i) = \sum_{i=1}^m \theta_i. \end{aligned}$$

Leonard (1976) shows that in the limit $m \to \infty$, the risk of the estimator (2.32) with respect to SEL_0 is approximately given by

$$\left\{\frac{S(\boldsymbol{\theta})}{S(\boldsymbol{\theta}) + \bar{\theta}}\right\} \sum_{i=1}^m \theta_i,$$

where $\bar{\theta} = \sum_{i=1}^m \theta_i/m$ and $S(\boldsymbol{\theta}) = \sum_{i=1}^m (\theta_i - \bar{\theta})^2/(m-1)$. Thus, the above expression means that as m approaches infinity, the risk of Leonard's estimator $\delta_i^L(\mathbf{Y})$ is smaller than that of the MLE by an approximate multiplicative factor of $\frac{S(\boldsymbol{\theta})}{S(\boldsymbol{\theta})+\bar{\theta}}$.

For the same problem, Morris (1983b) suggests an EB method which is similar to Leonard's estimator in (2.32), with a slight modification for the shrinking weight. His estimator is given as

$$\delta_i^M(\mathbf{Y}) = (1 - C_M)\ Y_i + C_M\ \overline{Y} \tag{2.34}$$

with

$$C_M = \frac{m-3}{m-1} \frac{\overline{Y}}{(s^2 - \overline{Y})_+ + \overline{Y}} \tag{2.35}$$

and s^2 as for (2.33). In fact, the second part in the right hand side of (2.35) is the same as the shrinkage proportion of C_L in (2.33). However, the multiplicative adjustment $\frac{m-3}{m-1}$ serves as a correction to the fact that EB estimation ignores the variation in the parameters that are estimated from the data, and therefore results in more radical shrinking. With the same objective, Kass and Steffey (1989) suggest Laplacian approximations leading to estimates for the posterior variance of the Poisson means which account for the uncertainty involved with the hyperparameters.

Albert type estimators

Albert (1981) considers a conjugate Poisson/Gamma situation and derives a shrinkage estimator that smoothes towards the prior means of the Poisson parameters. His estimator is similar to the posterior mean, which in this case takes the linear form (2.29), and therefore improves significantly over the MLE in the region of the prior mean. Furthermore, this method performs better than the posterior mean when the prior information is misleading, by restricting the amount of shrinking in this case.

When the sampling distribution (2.27) is Poison(θ_i), and the prior distribution of θ_i, $i = 1, \ldots, m$, in (2.28) is assumed to be gamma with parameters α_i and β_i respectively, Albert's estimator takes a form which is similar to the Efron and Morris type, and is given as

$$\delta_i^A(\mathbf{Y}) = (1 - C_A)\, Y_i + C_A\, M_i,$$

where M_i is the prior expectation $\frac{\alpha_i}{\beta_i}$ and C_A is the quantity minimising the risk of $\delta_i^A(\mathbf{Y})$ under SEL_0, with the involved unknown parameters θ_i, $i = 1, \ldots, m$, substituted by their MLE Y_i, that is

$$C_A = \min\left\{1, C_A(\mathbf{Y})\, \frac{\beta_i}{\beta_i + 1}\right\},$$

with

$$C_A(\mathbf{Y}) = \frac{\sum_{j=1}^m Y_j \left(\frac{\beta_j}{\beta_j+1}\right)}{\sum_{j=1}^m \left\{\frac{(Y_j - M_j)\beta_j}{\beta_j+1}\right\}^2 + \sum_{j=1}^m Y_j \left(\frac{\beta_j}{\beta_j+1}\right)^2}.$$

The requirement that C_A does not exceed one prevents the estimator from shrinking beyond the prior mean M_i, thus making the estimator behave like the Bayes

rule for the conjugate Poisson/Gamma model. However, the role of $C_A(\mathbf{Y})$ is to restrict the shrinkage of the EB estimator when the data are not close to the prior mean, leading to an estimator with better risk properties than the Bayes rule under prior misspecification. To use the estimator, the hyperparameters M_i and β_i can then be substituted by their method of moments estimates. However, assuming that the Poisson parameters $\theta_i,\ i = 1, \ldots, m$, come from a common gamma distribution with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$ we obtain the EB estimator

$$\delta_i^A(\mathbf{Y}) = (1 - C_A)\ Y_i + C_A\ \overline{Y} \tag{2.36}$$

where the shrinking coefficient is given by

$$C_A = \frac{m\overline{Y}}{(m-1)s^2 + m\overline{Y}}. \tag{2.37}$$

This is the same as an estimator derived by Hudson (1974), and we will refer to it as the Albert type estimator.

2.3.2 Hierarchical Bayes

In a full Bayesian approach, the first stage hyperparameters, that is the parameters of the prior distribution of $\theta_i,\ i = 1, \ldots, m$, are treated as random variables, rather than estimated from the data. Therefore, the idea of shrinkage estimation can be extended to hierarchical modelling, by assuming one or more further levels of prior information for the hyperparameters. Albert (1985) and Leonard and Novick (1986) consider approximate shrinkage estimators for the conjugate hierarchical Poisson/Gamma model. Applications of such methods in one and two-way contingency tables can be also found in Albert (1988), while Christiansen and Morris (1997) suggest approximations which extend these ideas to hierarchical Poisson regression. Most of the recent work on the estimation for hierarchical Poisson models has exploited the advance of simulation based methods for Bayesian inference, and mainly the Markov chain Monte Carlo methodology. The inference problem in the hierarchical model will be examined separately in subsequent chapters.

2.4 Summary and conclusions

In the present chapter we have reviewed some of the methods that have been proposed for tackling the problem of the simultaneous estimation of the means of several conditionally independent Poisson distributions using linear rules.

The inadmissibility of the MLE in this situation has motivated an effort to derive estimators with improved risk properties. The multiparameter nature of the

problem offers the possibility of exploiting information from all the involved parameters, thus leading to the so-called shrinkage linear estimators. These smooth the MLE towards a fixed, data-determined or prior-guess point, yielding improved estimates under a specific loss function. Various authors obtain such estimators constructing a rule that shrinks towards a chosen point in a way such that the resulting estimator universally dominates the MLE in terms of risk under the loss function of their choice. Other authors argue that one should only be interested in a limited region of the parameter space, that being the region within which $\boldsymbol{\theta}$ is likely to lie, and derive their estimators accordingly.

The methodology reviewed in this chapter demonstrates that we can obtain estimators that have smaller risk than the MLE when several conditionally independent Poisson means are to be estimated simultaneously. However, in order to choose among the suggested methods, one should take into consideration that each rule is constructed so that it improves the risk of the MLE under a given loss function and it shrinks towards a possibly different point. The adoption of a unique loss function is usually a difficult task, depending on the nature of the problem for which a decision must be made. The choice of the direction of shrinkage may also require some knowledge of the distributional characteristics of the problem under consideration, and can also demand appropriate use of prior information. In any case, if good risk properties is the primary aim, one should examine the performance of the considered methods in a wide range of the parameter space and under a variety of different loss functions.

Chapter 3

Empirical Bayes estimation for a Poisson/log-normal model

3.1 Introduction

In Chapter 2 we reviewed various methods for simultaneously estimating several Poisson means. We initially considered, according to model (1.1), the case where no prior information was available and then, in (2.27) and (2.28), we assumed a prior mean and variance for the Poisson means θ_i, $i = 1, \ldots, m$, without any further distributional assumptions. The case where a conjugate gamma prior distribution is specified was also mentioned. Conjugate models are mathematically tractable as they provide closed forms for the posterior densities of the parameters of interest. Nevertheless, they do not always offer the possibility of flexible statistical modelling as they restrict the choice of prior and, in some cases, hyperprior specifications. In the Poisson case, assuming a conjugate gamma prior for the parameters θ_i, $i = 1, \ldots, m$, does not allow the assumption of certain forms of noninformative hyperpriors for the first stage parameters, as these may lead to improper posterior distributions. Using such noninformative priors would be a reasonable choice under the assumption that little on no prior knowledge is available as we move to lower stages of the hierarchy.

We introduce an alternative formulation, assuming a log-normal prior distribution for the Poisson means. This prior setting might also be preferred when we wish to allow for more dispersion in the parameters θ_i, $i = 1, \ldots, m$, and possibly assume a correlation structure, which cannot be accommodated using the conjugate prior distribution. The Poisson/log-normal model has earlier been examined by Gaver and O'Muircheartaigh (1987), Carlin and Gelfand (1991), Tierney (1994), Christiansen and Morris (1997) and others. Clayton and Kaldor (1987) employ the aspect of shrinkage estimation in a Poisson/log-normal formulation in a small area estimation case, and Breslow and Clayton (1993) consider

the same problem under the more general context of random effects models.

3.1.1 The model formulation

The Bayesian approach that we adopt requires that a prior stage should be added in model (1.1). Thus, given the parameters $\theta_1, \theta_2, \ldots, \theta_m$, we assume that $Y_1, Y_2, \ldots, Y_m$, are independent Poisson random variables with means $\theta_1, \theta_2, \ldots, \theta_m$, respectively. Furthermore, for the specification of the prior structure, we assume that the logarithms of the Poisson means, that is $\gamma_i = \log(\theta_i), i = 1, \ldots, m$, are independently and normally distributed with a common mean μ and a common variance σ^2. This is equivalent to assuming that the Poisson parameters θ_i, $i = 1, \ldots, m$, follow independent and identical log-normal distributions, with parameters μ and σ^2. The model can be expressed as following:

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \theta_i &\overset{iid}{\sim} \text{LN}(\mu, \sigma^2), \end{aligned} \tag{3.1}$$

where $i = 1, \ldots, m$. From the theory of the log-normal distribution (e.g. Aitchison and Brown, 1957) the mean and variance of θ_i, denoted by ξ and ϕ respectively, are given by

$$\begin{aligned} \xi &= \text{E}(\theta_i) = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \\ \phi &= \text{var}(\theta_i) = \xi^2\left(\text{e}^{\sigma^2} - 1\right). \end{aligned} \tag{3.2}$$

3.2 Posterior inference for the Poisson/log-normal model

In model (3.1) the hierarchy stops at the first stage of prior specification, meaning that the hyperparameters μ and σ^2 will only enter the analysis either as known constants or as empirically estimated quantities. However, even when preliminary analyses have been conducted, the assignment of a fixed value to the hyperparameters based on past experience, is not an easy task. Alternatively, their evaluation using the current data leads to the adoption of an EB approach, which will be the subject of a subsequent section. Once μ and σ^2 have been evaluated, we can use the posterior distribution $p(\theta_i|\mathbf{y}, \mu, \sigma^2)$ to draw inferences about the Poisson means θ_i, $i = 1, \ldots, m$. Employing the parametrisation $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_m)^{\text{T}}$ with $\gamma_i = \log(\theta_i)$, we let $L(\boldsymbol{\gamma}|\mathbf{y})$ and $\pi(\boldsymbol{\gamma}|\mu, \sigma^2)$ denote the likelihood and the prior distribution of the parameter vector $\boldsymbol{\gamma}$ respectively. Then, according to the

model structure (3.1), these are given by

$$L(\boldsymbol{\gamma}|\mathbf{y}) = \prod_{i=1}^{m} \left\{ \frac{\exp(\gamma_i y_i - e^{\gamma_i})}{y_i!} \right\}$$

and

$$\pi(\boldsymbol{\gamma}|\mu, \sigma^2) = \prod_{i=1}^{m} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2 \right\} \right].$$

Therefore, applying Bayes' theorem, the posterior density of the parameter vector $\boldsymbol{\gamma}$ is given by

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{y}) &\propto L(\boldsymbol{\gamma}|\mathbf{y})\, \pi(\boldsymbol{\gamma}|\mu, \sigma^2) \\ &\propto \exp\left[\sum_{i=1}^{m} \left\{ (\gamma_i y_i - e^{\gamma_i}) - \frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2 \right\} \right], \end{aligned} \tag{3.3}$$

where in the notation for the posterior density we suppress the dependence upon μ and σ^2 since they are treated as constants. The conditional independence of the γ_i, $i = 1, \ldots, m$, implies that the posterior density of each component of the vector $\boldsymbol{\gamma}$ can be written as

$$p(\gamma_i|\mathbf{y}) \propto \exp\{(\gamma_i y_i - e^{\gamma_i}) - \frac{1}{2}\,\sigma^{-2}(\gamma_i - \mu)^2\}, \quad i = 1, \ldots, m. \tag{3.4}$$

The following definitions introduce the concepts of average risk and Bayes estimator, which will be used throughout this thesis.

Definition 3.1. *Let $\boldsymbol{\delta}(\mathbf{Y})$ denote an estimator of the parameter vector $\boldsymbol{\theta}$ and $R\{\boldsymbol{\delta}(\mathbf{Y})\}$ be the risk associated with a loss function $L\{\boldsymbol{\delta}(\mathbf{Y}), \boldsymbol{\theta}\}$, that is $R\{\boldsymbol{\delta}(\mathbf{Y})\} = E_{Y|\theta}[L\{\boldsymbol{\delta}(\mathbf{Y}), \boldsymbol{\theta}\}]$. Also, let $\pi(\cdot)$ be a prior distribution over the space of $\boldsymbol{\theta}$. Then, the average risk of the estimator $\boldsymbol{\delta}(\mathbf{Y})$ with respect to $\pi(\cdot)$, denoted by $R_\pi\{\boldsymbol{\delta}(\mathbf{Y})\}$, is defined as (e.g. see Carlin and Louis, 1996)*

$$R_\pi\{\boldsymbol{\delta}(\mathbf{Y})\} = E_\theta R\{\boldsymbol{\delta}(\mathbf{Y})\} = E_\theta E_{Y|\theta}[L\{\boldsymbol{\delta}(\mathbf{Y}), \boldsymbol{\theta}\}],$$

where the first expectation is taken with respect to the prior distribution $\pi(\cdot)$ of θ.

Definition 3.2. *An estimator $\boldsymbol{\delta}(\mathbf{Y})$ of $\boldsymbol{\theta}$ is a Bayes estimator with respect to the prior distribution $\pi(\boldsymbol{\theta})$, if it minimises the average risk $R_\pi\{\boldsymbol{\delta}(\mathbf{Y})\}$.*

Inferences about the Poisson parameters θ_i, $i = 1, \ldots, m$, may be based on the posterior density $p(\theta_i|\mathbf{y})$, obtained with an appropriate transformation of the posterior density of γ_i in (3.4). We can estimate θ_i, $i = 1, \ldots, m$, employing the posterior mean, which is the Bayes estimator under the squared error loss function. According to Definition 3.2, this means that the posterior mean minimises

the average risk under the squared error loss function. The posterior mean of $\theta_i = e^{\gamma_i}$ is

$$\mathrm{E}(\theta_i|\mathbf{y}) = \frac{\int e^{\gamma_i}\, p(\gamma_i|\mathbf{y})\, d\gamma_i}{\int p(\gamma_i|\mathbf{y})\, d\gamma_i}, \qquad i = 1, \ldots, m, \tag{3.5}$$

where the posterior density $p(\gamma_i|\mathbf{y})$ is given in (3.4) and the denominator provides for the normalising constant.

As a result of our assumption for the prior distribution of the Poisson means, the posterior density $p(\gamma_i|\mathbf{y})$ does not take a closed analytical form and the integrations needed in the computation of the posterior expectation in (3.5) do not have an analytical solution. This mathematical intractability is a common problem with Bayesian analysis when nonconjugate models are considered. A possible way to overcome this problem is to obtain approximations, either to the posterior distribution, or to the posterior mean directly. Two approximations to the posterior mean are derived in the remaining of this chapter.

Other measures of location can be also used as estimates of the Poisson means θ_i, $i = 1, \ldots, m$. For instance, the median of the posterior distribution may be employed, this being the Bayes estimator under the absolute error loss function. Also, various percentiles of the posterior distribution can be used to obtain interval estimates. However, once again the nonclosed form of the posterior distribution implies that all these distributional characteristics cannot be derived analytically, nor can they be obtained with direct simulation.

3.3 A linear approximation to the posterior mean

In Section 2.3 the posterior mean was presented as a linear estimator of θ_i, taking the form $\mathrm{E}(\theta_i|\mathbf{y}) = aY_i + b$. When the posterior mean can be given exactly in this form, that is when a conjugate prior distribution is assumed, the quantities a and b are explicitly expressed in terms of the prior moments, as given in (2.30). Although the Poisson/log-normal model does not allow this linear expression for the expectation of the posterior distribution, we can use a linear combination of functions of the data as an approximation. Then, the problem to be answered is how to determine the coefficients of such a linear estimator.

3.3.1 General derivation of the best linear predictor

We now describe a general derivation of a linear approximation to the posterior mean. Let $\mathbf{Y} = (Y_1, \ldots, Y_m)^{\mathrm{T}}$ denote a vector of observable random variables and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}}$ the vector consisting of the parameters of interest. We

assume that given $\boldsymbol{\theta}$, $\mathbf{Y}$ has a joint distribution with population mean vector $\mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\theta})$ and population covariance matrix $\mathrm{cov}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}) = \mathbf{V}(\boldsymbol{\theta})$. We also assume a prior distribution for the parameter vector $\boldsymbol{\theta}$ and we let the prior expectation vector be $\mathrm{E}(\boldsymbol{\theta}) = \boldsymbol{\theta}^*$ and the prior covariance matrix $\mathrm{cov}(\boldsymbol{\theta}) = \boldsymbol{\Lambda}$. Furthermore, we denote the prior expectation vector of the population mean vector by $\mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{\mu}(\boldsymbol{\theta})\} = \boldsymbol{\mu}^*$, its prior covariance matrix by $\mathrm{cov}_{\boldsymbol{\theta}}\{\boldsymbol{\mu}(\boldsymbol{\theta})\} = \mathbf{D}$ and the prior expectation matrix of the population covariance matrix by $\mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{cov}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y})\} = \mathbf{C}$.

We want to derive an approximation to the posterior mean of the parameter vector $\boldsymbol{\theta}$, in such a way that it is the best linear predictor (BLP) for $\boldsymbol{\theta}$ in the sense that it is a linear function of the data and satisfies the property of minimum average risk. The latter implies that the BLP will be the closest linear approximation to the posterior mean in terms of expected distance from the true value $\boldsymbol{\theta}$, with respect to both the prior and the data distribution. Suppose that $\widetilde{\boldsymbol{\theta}}$ is a linear approximation to the posterior mean of $\boldsymbol{\theta}$, having the form

$$\widetilde{\boldsymbol{\theta}} = \mathbf{a} + \mathbf{A}\mathbf{Y},$$

where the $m \times 1$ vector $\mathbf{a}$ and the $m \times m$ matrix $\mathbf{A}$ are to be defined. We specify the summed quadratic loss function for $\widetilde{\boldsymbol{\theta}}$, given by

$$L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\mathrm{T}}(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}).$$

From Definition 3.1, the average risk is given as $R_\pi(\widetilde{\boldsymbol{\theta}}) = \mathrm{E}_{\boldsymbol{\theta}}\mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\{L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})\} = \mathrm{E}_{\boldsymbol{\theta}}\{R(\widetilde{\boldsymbol{\theta}})\}$, where $R(\widetilde{\boldsymbol{\theta}})$ denotes the frequentist risk of $\widetilde{\boldsymbol{\theta}}$, given in Definition 2.1. Under the summed quadratic loss function, $R(\widetilde{\boldsymbol{\theta}})$ is the MSE and can be written as

$$\begin{aligned} R(\widetilde{\boldsymbol{\theta}}) &= \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left\{L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})\right\} \\ &= \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left\{(\mathbf{a} + \mathbf{A}\mathbf{Y} - \boldsymbol{\theta})^{\mathrm{T}}(\mathbf{a} + \mathbf{A}\mathbf{Y} - \boldsymbol{\theta})\right\} \\ &= \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left\{\mathbf{a}^{\mathrm{T}}(\mathbf{a} + \mathbf{A}\mathbf{Y} - \boldsymbol{\theta}) + \mathbf{Y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{a} + \mathbf{A}\mathbf{Y} - \boldsymbol{\theta}) - \boldsymbol{\theta}^{\mathrm{T}}(\mathbf{a} + \mathbf{A}\mathbf{Y} - \boldsymbol{\theta})\right\}. \end{aligned}$$

We can now use the first conditional moment of $\mathbf{Y}$, that is $\mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\theta})$, to obtain

$$\begin{aligned} R(\widetilde{\boldsymbol{\theta}}) &= \mathbf{a}^{\mathrm{T}}\{\mathbf{a} + \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\} + \boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\mathbf{a} + \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{Y}\right) \\ &\quad -\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{T}}\{\mathbf{a} + \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\} \\ &= \mathbf{a}^{\mathrm{T}}\{\mathbf{a} + \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\} - \boldsymbol{\theta}^{\mathrm{T}}\{\mathbf{a} + \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\} \\ &\quad +\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}(\mathbf{a} - \boldsymbol{\theta}) + \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{Y}\right), \end{aligned}$$

and completing the square for $\{\mathbf{a} + \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\}$ we obtain

$$\begin{aligned} R(\widetilde{\boldsymbol{\theta}}) &= \{\mathbf{a} + \mathbf{A}\,\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\}^{\mathrm{T}}\{\mathbf{a} + \mathbf{A}\,\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\} \\ &\quad -\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) + \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{Y}\right). \end{aligned} \tag{3.6}$$

Then, using that

$$\begin{aligned} \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right) &= \mathrm{cov}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}) + \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y})\,\mathrm{E}^{\mathrm{T}}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}) \\ &= \mathbf{V}(\boldsymbol{\theta}) + \boldsymbol{\mu}(\boldsymbol{\theta})\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta}), \end{aligned}$$

and from the properties of the trace of matrices, the expectation $\mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{Y}\right)$ in (3.6) is given by

$$\begin{aligned} \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{Y}\right) &= \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left\{\mathrm{tr}\left(\mathbf{Y}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{Y}\right)\right\} = \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left\{\mathrm{tr}\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)\right\} \\ &= \mathrm{tr}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\left(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\right)\right\} = \mathrm{tr}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{V}(\boldsymbol{\theta}) + \mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta})\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\right\} \\ &= \mathrm{tr}\left\{\mathbf{A}\mathbf{V}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\right\} + \boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}). \end{aligned}$$

Therefore, the frequentist risk in (3.6) can be written as

$$R(\widetilde{\boldsymbol{\theta}}) = \left\{\mathbf{a} + \mathbf{A}\,\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\right\}^{\mathrm{T}}\left\{\mathbf{a} + \mathbf{A}\,\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\right\} + \mathrm{tr}\left\{\mathbf{A}\,\mathbf{V}(\boldsymbol{\theta})\,\mathbf{A}^{\mathrm{T}}\right\}.$$

It follows that the average risk is given by

$$\begin{aligned} R_{\pi}(\widetilde{\boldsymbol{\theta}}) &= \mathrm{E}_{\boldsymbol{\theta}}\mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}\{L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})\} = \mathrm{E}_{\boldsymbol{\theta}}\{R(\widetilde{\boldsymbol{\theta}})\} \\ &= \mathrm{E}_{\boldsymbol{\theta}}[\{\mathbf{a} + \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\}^{\mathrm{T}}\{\mathbf{a} + \mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta}) - \boldsymbol{\theta}\}] + \mathrm{tr}\left[\mathrm{E}_{\boldsymbol{\theta}}\left\{\mathbf{A}\mathbf{V}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\right\}\right]. \end{aligned}$$

Taking into account the prior expectations $\mathrm{E}(\boldsymbol{\theta}) = \boldsymbol{\theta}^*$, $\mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{\mu}(\boldsymbol{\theta})\} = \boldsymbol{\mu}^*$ and $\mathrm{E}_{\boldsymbol{\theta}}\{\mathbf{V}(\boldsymbol{\theta})\} = \mathbf{C}$, the above equation gives

$$\begin{aligned} R_{\pi}(\widetilde{\boldsymbol{\theta}}) &= \mathbf{a}^{\mathrm{T}}\left(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}^* - \boldsymbol{\theta}^*\right) + \boldsymbol{\mu}^{*\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{a} + \mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta})\right\} \\ &\quad - \mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}\right\} - \boldsymbol{\theta}^{*\mathrm{T}}\mathbf{a} - \mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{A}\boldsymbol{\mu}(\boldsymbol{\theta})\right\} + \mathrm{E}_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}\right) + \mathrm{tr}\left\{\mathbf{A}\mathbf{C}\mathbf{A}^{\mathrm{T}}\right\}, \end{aligned}$$

and using that

$$\begin{aligned} \boldsymbol{\Lambda} = \mathrm{cov}(\boldsymbol{\theta}) &= \mathrm{E}\left(\boldsymbol{\theta}\boldsymbol{\theta}^{\mathrm{T}}\right) - \boldsymbol{\theta}^*\boldsymbol{\theta}^{*\mathrm{T}} \\ \mathbf{D} = \mathrm{cov}_{\boldsymbol{\theta}}\{\boldsymbol{\mu}(\boldsymbol{\theta})\} &= \mathrm{E}\left\{\boldsymbol{\mu}(\boldsymbol{\theta})\boldsymbol{\mu}^{\mathrm{T}}(\boldsymbol{\theta})\right\} - \boldsymbol{\mu}^*\boldsymbol{\mu}^{*\mathrm{T}}, \end{aligned}$$

together with the trace properties, we have

$$\begin{aligned} R_{\pi}(\widetilde{\boldsymbol{\theta}}) &= \mathbf{a}^{\mathrm{T}}\left(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}^* - \boldsymbol{\theta}^*\right) + \boldsymbol{\mu}^{*\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{a} + \mathrm{tr}\left\{\mathbf{A}^{\mathrm{T}}\mathbf{A}\left(\mathbf{D} + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*\mathrm{T}}\right)\right\} \\ &\quad - 2\,\mathrm{tr}\left[\mathbf{A}\mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\mu}(\boldsymbol{\theta})\boldsymbol{\theta}^{\mathrm{T}}\right\}\right] - \boldsymbol{\theta}^{*\mathrm{T}}\mathbf{a} + \mathrm{tr}\left\{\boldsymbol{\Lambda} + \boldsymbol{\theta}^{*\mathrm{T}}\boldsymbol{\theta}^*\right\} + \mathrm{tr}\left\{\mathbf{A}\mathbf{C}\mathbf{A}^{\mathrm{T}}\right\}. \end{aligned}$$

Then, completing the square for $(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}^* - \boldsymbol{\theta}^*)$, we obtain

$$\begin{aligned} R_{\pi}(\widetilde{\boldsymbol{\theta}}) &= \left(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}^* - \boldsymbol{\theta}^*\right)^{\mathrm{T}}\left(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}^* - \boldsymbol{\theta}^*\right) + \boldsymbol{\mu}^{*\mathrm{T}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\theta}^* + \boldsymbol{\theta}^{*\mathrm{T}}\mathbf{A}\boldsymbol{\mu}^* \\ &\quad + \mathrm{tr}\left[\mathbf{A}\mathbf{D}\mathbf{A}^{\mathrm{T}} - 2\mathbf{A}\,\mathrm{E}_{\boldsymbol{\theta}}\left\{\boldsymbol{\mu}(\boldsymbol{\theta})\boldsymbol{\theta}^{\mathrm{T}}\right\} + \boldsymbol{\Lambda}\right] + \mathrm{tr}\left\{\mathbf{A}\mathbf{C}\mathbf{A}^{\mathrm{T}}\right\} \\ &= \left(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}^* - \boldsymbol{\theta}^*\right)^{\mathrm{T}}\left(\mathbf{a} + \mathbf{A}\boldsymbol{\mu}^* - \boldsymbol{\theta}^*\right) + 2\,\mathrm{tr}\left(\boldsymbol{\mu}^*\boldsymbol{\theta}^{*\mathrm{T}}\mathbf{A}\right) \\ &\quad + \mathrm{tr}\left\{\mathbf{A}\mathbf{D}\mathbf{A}^{\mathrm{T}} - 2\mathbf{A}\left(\mathbf{G} + \boldsymbol{\mu}^*\boldsymbol{\theta}^{*\mathrm{T}}\right) + \boldsymbol{\Lambda} + \mathbf{A}\mathbf{C}\mathbf{A}^{\mathrm{T}}\right\} \end{aligned} \tag{3.7}$$

where

$$\begin{aligned}\mathbf{G} &= \mathrm{E}\left\{\boldsymbol{\mu}(\boldsymbol{\theta})\,\boldsymbol{\theta}^{\mathrm{T}}\right\} - \boldsymbol{\mu}^{*}\,\boldsymbol{\theta}^{*\mathrm{T}} \\ &= \mathrm{cov}\left\{\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\theta}^{\mathrm{T}}\right\}. \end{aligned} \tag{3.8}$$

Finally, (3.7) implies that the average risk is given by

$$\begin{aligned} R_{\pi}(\widetilde{\boldsymbol{\theta}}) = &\left(\mathbf{a} + \mathbf{A}\,\boldsymbol{\mu}^{*} - \boldsymbol{\theta}^{*}\right)^{\mathrm{T}}\left(\mathbf{a} + \mathbf{A}\,\boldsymbol{\mu}^{*} - \boldsymbol{\theta}^{*}\right) \\ &+ \mathrm{tr}\left\{\boldsymbol{\Lambda} + \mathbf{A}\,(\mathbf{D} + \mathbf{C})\,\mathbf{A}^{\mathrm{T}} - 2\mathbf{A}\,\mathbf{G}\right\}. \end{aligned} \tag{3.9}$$

To minimise the average risk we first notice that the first term of the right hand side of (3.9) cannot be negative, and therefore it is minimised when it is equal to zero. This will give

$$\hat{\mathbf{a}} = \boldsymbol{\theta}^{*} - \mathbf{A}\,\boldsymbol{\mu}^{*}. \tag{3.10}$$

We then find $\mathbf{A}$ to satisfy $\frac{\partial}{\partial \mathbf{A}}\left[\mathrm{tr}\{\boldsymbol{\Lambda} + \mathbf{A}\,(\mathbf{D} + \mathbf{C})\,\mathbf{A}^{\mathrm{T}} - 2\mathbf{A}\,\mathbf{G}\}\right] = 0$. Using the matrix differentiation properties in Appendix C, and since the matrices $\mathbf{D}$, $\mathbf{C}$ and $\mathbf{G}$ are symmetric, we obtain

$$\begin{aligned} & 2\,(\mathbf{D} + \mathbf{C})\mathbf{A} - 2\,\mathbf{G} = 0 \\ \Rightarrow \quad & \widehat{\mathbf{A}} = (\mathbf{D} + \mathbf{C})^{-1}\,\mathbf{G}. \end{aligned} \tag{3.11}$$

Therefore the best linear predictor for $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \hat{\mathbf{a}} + \widehat{\mathbf{A}}\,\mathbf{Y} \tag{3.12}$$

with $\hat{\mathbf{a}}$ and $\widehat{\mathbf{A}}$ given in (3.10) and (3.11) respectively.

3.3.2 BLP in the Poisson/log-normal case

We will use the BLP as an estimator for the posterior mean of the parameters θ_i, $i = 1, \ldots, m$, in the Poisson/log-normal model. As stressed earlier, when a conjugate model structure is involved, the Bayes estimator under the quadratic loss function, that is the estimator with the smallest average risk with respect to the associated prior distribution, can be exactly expressed in a linear form, that being the posterior expectation of the parameters of interest. Therefore, when the nonconjugate Poisson/log-normal structure is assumed, the motivation is to obtain a linear approximation to the posterior mean which will again possess the property of minimum average risk within the class of all linear estimators of the same type, that is the BLP.

We can derive the BLP working from first principles and following the procedure described in the preceding subsection. Adopting a component-wise formulation, we let $\tilde{\theta}_i$ be a linear estimator of θ_i, having the form

$$\tilde{\theta}_i = a + AY_i.$$

Then, under the summed squared error loss function

$$L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \left(\tilde{\theta}_i - \theta_i\right)^2,$$

the average risk is given by

$$R_\pi(\tilde{\theta}_i) = \mathrm{E}_\theta \mathrm{E}_{Y|\theta} \left\{ \sum_{i=1}^{m} (a + AY_i - \theta_i)^2 \right\}.$$

Using the prior moments of θ_i given in (3.2), together with the moments of Y conditional on θ, i.e.

$$\mathrm{E}_{Y|\theta}(Y_i) = \mathrm{var}_{Y|\theta}(Y_i) = \theta_i, \qquad i = 1, \ldots, m,$$

we deduce that the average risk of $\tilde{\theta}_i$ under the summed squared error loss function is

$$R_\pi(\tilde{\theta}_i) = \sum_{i=1}^{m} \left[\{a - \xi(1 - A)\}^2 + (1 - A)^2\phi + A^2\xi \right].$$

We minimise the average risk in the above expression by setting

$$\hat{a} = (1 - A)\xi$$

and solving the equation $\frac{\partial}{\partial A} \{(1 - A)^2\phi + A^2\xi\} = 0$. This will give

$$\widehat{A} = \frac{\phi}{\phi + \xi}.$$

It follows that the BLP for the Poisson means θ_i, $i = 1, \ldots, m$, i.e. the linear estimator that has the smallest average risk under the summed squared error loss function is given as

$$\tilde{\theta}_i^{\mathrm{BLP}} = \frac{\phi}{\phi + \xi} Y_i + \frac{\xi}{\phi + \xi} \xi \tag{3.13}$$

where $\xi = \mathrm{E}(\theta_i)$ and $\phi = \mathrm{var}(\theta_i)$.

We can also derive the BLP for the parameters θ_i, $i = 1, \ldots, m$, in the Poisson/log-normal model applying the general result (3.12) directly. We let $\mathbf{e}_m$ denote the unit $m \times 1$ vector, $\mathbf{I}_m$ the $m \times m$ identity matrix, and $\mathrm{diag}(\boldsymbol{\theta})$ the

$m \times m$ matrix which has the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\mathrm{T}}$ on its diagonal and zero elsewhere. We then notice that the specification of the Poisson/log-normal model in (3.1) implies that, according to the notation introduced in Subsection 3.3.1, the conditional mean vector and covariance matrix of $\mathbf{Y}$ given $\boldsymbol{\theta}$ are given as

$$\begin{aligned}\boldsymbol{\mu}(\boldsymbol{\theta}) &= \mathrm{E}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}) = \boldsymbol{\theta} \\ \mathbf{V}(\boldsymbol{\theta}) &= \mathrm{cov}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y}) = \mathrm{diag}(\boldsymbol{\theta}),\end{aligned}$$

and the prior mean and covariance of the parameter vector $\boldsymbol{\theta}$ are

$$\begin{aligned}\boldsymbol{\theta}^* &= \mathrm{E}(\boldsymbol{\theta}) = \xi \mathbf{e}_m \\ \boldsymbol{\Lambda} &= \mathrm{cov}(\boldsymbol{\theta}) = \phi \mathbf{I}_m.\end{aligned}$$

Furthermore, the expectation and the covariance of the mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ with respect to the prior distribution of $\boldsymbol{\theta}$, are given by

$$\begin{aligned}\boldsymbol{\mu}^* &= \mathrm{E}_{\boldsymbol{\theta}}\{\boldsymbol{\mu}(\boldsymbol{\theta})\} = \mathrm{E}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \xi \mathbf{e}_m \\ \mathbf{D} &= \mathrm{cov}_{\boldsymbol{\theta}}\{\boldsymbol{\mu}(\boldsymbol{\theta})\} = \mathrm{cov}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \phi \mathbf{I}_m,\end{aligned}$$

and finally, the prior expectation of the conditional covariance of vector $\mathbf{Y}$ given $\boldsymbol{\theta}$, and the covariance matrix between vectors $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}$, can be respectively expressed as

$$\begin{aligned}\mathbf{C} = \mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{cov}_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{Y})\} = \mathrm{E}_{\boldsymbol{\theta}}\{\mathbf{V}(\boldsymbol{\theta})\} = \mathrm{E}_{\boldsymbol{\theta}}\{\mathrm{diag}(\boldsymbol{\theta})\} = \xi \mathbf{I}_m \\ \mathbf{G} = \mathrm{cov}\left\{\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\theta}^{\mathrm{T}}\right\} = \mathrm{cov}\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{T}}\right) = \phi \mathbf{I}_m.\end{aligned}$$

Then, using (3.12) we obtain the BLP as

$$\begin{aligned}\widetilde{\boldsymbol{\theta}}^{\mathrm{BLP}} &= \hat{\mathbf{a}} + \widehat{\mathbf{A}}\,\mathbf{Y} \\ &= \boldsymbol{\theta}^* - \left\{(\mathbf{D}+\mathbf{C})^{-1}\mathbf{G}\right\}\boldsymbol{\mu}^* + (\mathbf{D}+\mathbf{C})^{-1}\mathbf{G}\mathbf{Y} \\ &= \xi \mathbf{e}_m - \left\{(\phi \mathbf{I}_m + \xi \mathbf{I}_m)^{-1}\phi \mathbf{I}_m\right\}\xi \mathbf{e}_m + \left\{(\phi \mathbf{I}_m + \xi \mathbf{I}_m)^{-1}\phi \mathbf{I}_m\right\}\mathbf{Y} \\ &= \left(\frac{\phi}{\phi+\xi}\right)\mathbf{Y} + \left(\frac{\xi}{\phi+\xi}\right)\xi \mathbf{e}_m\end{aligned}$$

which component-wise gives the BLP for θ_i in (3.13). We notice here that this linear estimator of the posterior mean of θ_i can be written as

$$\tilde{\theta}_i^{\mathrm{BLP}} = (1-c)Y_i + c\mathrm{E}(\theta_i), \tag{3.14}$$

where the coefficient c is given by

$$c = \frac{\mathrm{E}(\theta_i)}{\mathrm{var}(\theta_i) + \mathrm{E}(\theta_i)}. \tag{3.15}$$

Equations (3.14) and (3.15) imply that our effort to obtain a linear estimator with minimum average risk, delivered a method which is the same as the conditional posterior mean (2.29) derived by Ericson (1969) and the Efron and Morris (1973) estimator in (2.31). The BLP of the Poisson means θ_i, $i = 1, \ldots, m$, is a shrinkage estimator, adjusting the usual estimate Y_i towards the prior mean of θ_i, with a weight which is inversely proportional to the prior variance ϕ, thus reflecting our confidence in the prior assumptions. When the prior variance is small, the shrinking coefficient tends to the value one and the estimator smoothes more towards the prior mean ξ. On the other hand, when our prior beliefs are rather vague, as this can be expressed by a large prior variance, the shrinking proportion tends to zero, attaching more weight to the MLE. One remaining problem with using the BLP to estimate the Poisson means, is the evaluation of the prior mean and variance of θ_i, $i = 1, \ldots, m$. We will return to this problem later in this chapter.

3.4 Importance sampling approximation to the posterior mean

The characteristics of a distribution which is not given in a known form can be estimated with the use of simulation methods. Inferences can be based on a sufficiently large number of values drawn from the distribution of interest. For instance, suitable functions of the simulated values can be averaged to give estimates for the mean or the standard deviation; ordered values may be used to estimate various percentiles; scatter plots and other graphs can be employed to examine the whole distribution, etc. However, simulation methods rely on the ability to sample the distribution under consideration. When this distribution does not appear in a known form, direct simulation is not possible and therefore alternative techniques must be adopted. We will describe a method that copes with such situations, namely the importance sampling technique. Following the problem introduced in the present chapter, we first give a general presentation of how importance sampling may be used to estimate the posterior mean of a function of a parameter of interest.

The general method

Let $g(\theta)$ be a function of a parameter θ. Here we take θ to be a scalar, but the same results hold for a vector parameter. We also let $p(\theta|\mathrm{y})$ denote the non-normalised posterior density of θ. Suppose we are interested in the posterior

expectation of $g(\theta)$. This can be written as

$$\mathrm{E}\{g(\theta)|\mathbf{y}\} = \frac{\int g(\theta)p(\theta|\mathbf{y})\,d\theta}{\int p(\theta|\mathbf{y})\,d\theta}$$

$$= \frac{\int g(\theta)\frac{p(\theta|\mathbf{y})}{I(\theta)}I(\theta)\,d\theta}{\int \frac{p(\theta|\mathbf{y})}{I(\theta)}I(\theta)\,d\theta} = \frac{\mathrm{E}_I\left\{g(\theta)\frac{p(\theta|\mathbf{y})}{I(\theta)}\right\}}{\mathrm{E}_I\left\{\frac{p(\theta|\mathbf{y})}{I(\theta)}\right\}}$$

where $I(\cdot)$ denotes the so-called importance density, which is the probability density function of a distribution from which we can simulate, and the expectations in the last expression are taken with respect to $I(\theta)$. Then, the importance sampling approximation to the posterior expectation is given by

$$\widehat{\mathrm{E}}\{g(\theta)|\mathbf{y}\} = \frac{\sum_{j=1}^{N} g(\theta_j)w_j}{\sum_{i=j}^{N} w_j}$$

where N is the number of Monte Carlo simulations, $w_j = \frac{p(\theta_j|\mathbf{y})}{I(\theta_j)}$, $j = 1, \ldots, N$, are the so-called importance weights, and θ_j, $j = 1, \ldots, N$, are random variates simulated from the importance density $I(\theta)$.

Under the conditions (Geweke, 1989) that the support of the importance density includes the support of the posterior density, that θ_j, $j = 1, \ldots, N$, are an independently and identically distributed sample from the importance density and that the expectation exists, the strong law of large numbers implies that $\widehat{\mathrm{E}}\{g(\theta)|\mathbf{y}\} \xrightarrow{a.s.} \mathrm{E}\{g(\theta)|\mathbf{y}\}$ as $N \to \infty$. The choice of the importance density is critical to the performance of the method. In general $I(\theta)$ should be chosen to mimic $p(\theta|\mathbf{y})$ as closely as possible, with tails that do not decay faster than those of the latter. We can therefore use importance sampling to estimate the posterior mean of a distribution, as long as we can find a suitable approximation to $p(\theta|\mathbf{y})$ and generate a sufficiently large number of values from this approximation.

3.4.1 An importance sampling estimator for the Poisson/log-normal model

We now obtain an estimator for the Poisson means θ_i, $i = 1, \ldots, m$, in the Poisson/log-normal model which approximates the posterior mean in (3.5), using the technique presented in the preceding paragraph. To derive the estimator, we first rewrite the posterior density of γ_i given in (3.4) adding γ_i to the first part of the exponent and subtracting it from the second. This gives

$$p(\gamma_i|\mathbf{y}) \propto \exp\left\{\gamma_i(y_i+1) - e^{\gamma_i} - \frac{1}{2}\,\sigma^{-2}(\gamma_i-\mu)^2 - \gamma_i\right\}$$
$$\propto I(\gamma_i)\,W(\gamma_i), \tag{3.16}$$

for $i = 1, \ldots, m$, where

$$\begin{aligned} I(\gamma_i) &= \exp\{\gamma_i(y_i + 1) - e^{\gamma_i}\} \\ W(\gamma_i) &= \exp\left\{-\frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2 - \gamma_i\right\}. \end{aligned} \tag{3.17}$$

Then, if a random variable Z_i has a gamma distribution with mean and variance equal to $y_i + 1$, denoted by $\mathrm{Ga}(y_i + 1, 1)$, the density $\pi(\gamma_i)$ of $\gamma_i = \log(Z_i)$ is given by the appropriate transformation of the density of Z_i, denoted by $p(Z_i)$, that is

$$\begin{aligned} \pi(\gamma_i) &\propto p(Z_i) \left|\frac{dZ_i}{d\gamma_i}\right| \\ &\propto e^{\gamma_i y_i} \exp(-e^{\gamma_i})\, e^{\gamma_i} \\ &= \exp\{\gamma_i(y_i + 1) - e^{\gamma_i}\} \\ &= I(\gamma_i). \end{aligned}$$

Therefore, $I(\gamma_i)$ is proportional to the probability density function of $\gamma_i = \log(Z_i)$ where $Z_i \sim \mathrm{Ga}(y_i + 1, 1)$. The expression for the posterior density of γ_i was modified in (3.16) in order to allow for the case of zero counts, that is the case when $y_i = 0$.

Using the above notation, the posterior mean of $\theta_i = e^{\gamma_i}$ can be written as

$$\begin{aligned} \mathrm{E}(\theta_i|\mathbf{y}) &= \frac{\int e^{\gamma_i}\, p(\gamma_i|\mathbf{y})\, d\gamma_i}{\int p(\gamma_i|\mathbf{y})\, d\gamma_i} = \frac{\int e^{\gamma_i} \frac{p(\gamma_i|\mathbf{y})}{I(\gamma_i)}\, I(\gamma_i)\, d\gamma_i}{\int \frac{p(\gamma_i|\mathbf{y})}{I(\gamma_i)}\, I(\gamma_i)\, d\gamma_i} \\ &= \frac{\int e^{\gamma_i} W(\gamma_i)\, I(\gamma_i)\, d\gamma_i}{\int W(\gamma_i)\, I(\gamma_i)\, d\gamma_i} = \frac{\mathrm{E}_I\{e^{\gamma_i} W(\gamma_i)\}}{\mathrm{E}_I\{W(\gamma_i)\}}. \end{aligned}$$

The expectations in the last expression are taken with respect to the normalised distribution whose probability density function is proportional to $I(\gamma_i)$. Hence, taking $I(\gamma_i)$ as the importance density and $W(\gamma_i)$ as the weight function, the above expression leads to the importance sampling approximation to the posterior expectation of θ_i, $i = 1, \ldots, m$, given by

$$\tilde{\theta}_i^{\mathrm{IS}} = \widehat{\mathrm{E}}(\theta_i|\mathbf{y}) = \frac{\sum_{j=1}^{N} e^{\gamma_{ij}}\, W(\gamma_{ij})}{\sum_{j=1}^{N} W(\gamma_{ij})}, \tag{3.18}$$

with $\gamma_{ij} = \log(Z_{ij})$ and Z_{ij}, $j = 1, \ldots, N$, being independent random variables from a $\mathrm{Ga}(y_i + 1, 1)$ distribution, N is the number of Monte Carlo simulations, and $W(\gamma_{ij})$ is the importance weight given as

$$W(\gamma_{ij}) = \exp\left\{-\frac{1}{2}\,\sigma^{-2}(\gamma_{ij} - \mu)^2 - \gamma_{ij}\right\}, \tag{3.19}$$

for $j = 1, \ldots, N$.

The computation of this estimator only requires the generation of a sufficiently large number of independent random variates from a gamma distribution and therefore it provides a fairly easy implemented alternative to numerical integration. However, we notice that in order to use the importance sampling estimator of θ_i, we first need to evaluate the hyperparameters μ and σ^2 involved in the importance weights (3.19). This is the same problem as evaluating the prior mean and variance of θ_i in the estimator given in (3.14). The following section deals with this issue.

3.5 Empirical Bayes estimation

In the analysis leading to both the BLP and the importance sampling estimator of θ_i, the hyperprior parameters μ and σ^2, or equivalently $\xi = \mathrm{E}(\theta_i)$ and $\phi = \mathrm{var}(\theta_i)$, are treated as known constants. The hierarchical full Bayesian formulation, which will be presented in Chapter 4, requires that these parameters are included in the analysis, after some hyperprior distributions have been assigned to them. In the present model one could evaluate these parameters according to prior beliefs. However, such prior beliefs may be difficult to obtain, especially when previous analyses are not available, and could also be subject to criticism due to possible subjectivity. Alternatively, we can use the information included in the observed data to estimate ξ and ϕ, or μ and σ^2, adopting an EB approach.

In the EB framework we use the marginal distribution of the data to substitute the unknown hyperparameters either by their maximum likelihood (ML) or by their moments estimates.

Method of moments estimation

We first find the marginal mean and variance of the data. The marginal mean of Y_i is

$$\begin{aligned} \mathrm{E}(Y_i) &= \mathrm{E}_\theta\{\mathrm{E}_{Y|\theta}(Y_i)\} = \mathrm{E}_\theta\{\theta_i\} \\ &= \exp\left(\mu + \frac{1}{2}\sigma^2\right) = \xi, \end{aligned} \tag{3.20}$$

and the marginal variance is given by

$$\begin{aligned} \mathrm{var}(Y_i) &= \mathrm{E}_\theta\{\mathrm{var}_{Y|\theta}(Y_i)\} + \mathrm{var}_\theta\{\mathrm{E}_{Y|\theta}(Y_i)\} \\ &= \mathrm{E}_\theta(\theta_i) + \mathrm{var}_\theta\,(\theta_i) \\ &= \xi + \phi = \xi + \xi^2\left(\mathrm{e}^{\sigma^2} - 1\right), \end{aligned} \tag{3.21}$$

where $\xi, \phi \geq 0$. If we let $\bar{y}$ denote the sample mean of the data and s^2 the sample variance, i.e. $s^2 = \sum_{i=1}^{m}(y_i - \bar{y})^2/(m-1)$, then ξ and ϕ may be estimated by equating the marginal moments to their sample estimates, that is solving the equations

$$\bar{y} = \xi \quad \text{and} \quad s^2 = \xi + \phi,$$

which give

$$\begin{aligned} \widehat{\xi} &= \bar{y} \\ \widehat{\phi} &= \left(s^2 - \bar{y}\right)_+, \end{aligned} \tag{3.22}$$

where $(\cdot)_+$ denotes the positive part function. Solving for μ and σ^2 we obtain

$$\begin{aligned} \widehat{\mu} &= \log(\bar{y}) - \frac{1}{2}\,\widehat{\sigma}^2 \\ \widehat{\sigma}^2 &= \log\left\{1 + \frac{(s^2 - \bar{y})_+}{\bar{y}^2}\right\}. \end{aligned} \tag{3.23}$$

Maximum Likelihood estimation

The marginal likelihood of the data for the Poisson/log-normal model is not given in a closed analytical form and therefore we will assume a normal approximation using the marginal moments (3.20) and (3.21), i.e. we will assume that approximately

$$Y_i \dot{\sim} N(\xi, \lambda), \qquad i = 1, \ldots, m,$$

where

$$\lambda = \xi + \phi, \quad \xi, \phi \geq 0, \quad \text{and} \quad \lambda \geq \xi.$$

Then the likelihood function, denoted here by L, is given by

$$L \propto \lambda^{-\frac{1}{2}m} \exp\{-\frac{1}{2}\lambda^{-1}\sum_{i=1}^{m}(y_i - \xi)^2\}$$

and is maximised for

$$\widehat{\xi} = \bar{y} \quad \text{and} \quad \widehat{\lambda} = \max\left\{\frac{S_{yy}}{m}, \bar{y}\right\},$$

and thus

$$\widehat{\phi} = \widehat{\lambda} - \widehat{\xi} = \left\{\frac{S_{yy}}{m} - \bar{y}\right\}_+,$$

where S_{yy} is the sum of squares corrected for the sample mean, i.e. $S_{yy} = \sum_{i=1}^{m}(y_i - \bar{y})^2$, and $(\cdot)_+$ denotes the positive part function. In the estimate for λ we take the maximum of S_{yy}/m and $\bar{y}$ following the restriction $\lambda = \xi + \phi \geq \xi$.

We will use the estimates obtained with the method of moments to evaluate the hyperparameters ξ and ϕ, or μ and σ^2, for the implementation of the BLP and

the importance sampling estimator respectively, thus following an EB approach. An EB solution for the Poisson/log-normal model is described by Gaver and O'-Muircheartaigh (1987) where numerical integration is employed for the estimation of the hyperparameters and the derivation of the posterior mean. Also, Clayton and Kaldor (1987) present an EB analysis for the same model using the EM algorithm for the hyperparameters and a multivariate normal approximation to the posterior distribution. The EB methodology in general is discussed in Morris (1983a). The Bayesian orientation of the EB approach has been criticised, since with the EB analysis we drop the hyperprior assumptions and instead we use the data to obtain information at that stage of the prior specification of the model. Also, by using the data to estimate the hyperparameters we underestimate the uncertainty associated with these parameters, which results in less variable estimators. For the BLP and the importance sampling estimator this would imply more radical shrinking towards the prior mean and narrower Bayesian intervals.

We finally notice that considering the method of moment estimates for ξ and ϕ, and substituting in (3.14) and (3.15) for the BLP of θ_i, yields

$$\tilde{\theta}_i^{\text{BLP}} = \left(\frac{s^2 - \overline{Y}}{s^2}\right)_+ Y_i + \min\left\{\frac{\overline{Y}}{s^2}, 1\right\}\overline{Y}, \tag{3.24}$$

$i = 1, \ldots, m$, which is also the Leonard (1976) shrinkage estimator in (2.32).

3.6 Example: Audit data

We illustrate the use of the linear predictor and the importance sampling estimator analysing two real data sets. In the first example the data, given in Table 3.1, concern the number of errors found in audit samples of 9 different accounts. The data set was first analysed in Matsumura and Tsui (1982). These authors consider various linear shrinkage methods to estimate the mean number of auditing errors.

Table 3.1: *Audit data (Matsumura and Tsui, 1982).*

Auditing errors y	0	1	2	3	6
Observed frequency	3	2	2	1	1

In our analysis, following the distributional assumptions (3.1), given the parameters θ_i, $i = 1, \ldots, 9$, each observation is modelled as a conditionally independent Poisson variable Y_i with corresponding means θ_i, $i = 1, \ldots, 9$, which are the

Table 3.2: *EB estimates of the expected number of errors* θ_i, $i = 1, \ldots, 9$, *in the audit data example.*

		EB estimates			
par.	y	*post. mean*	$\tilde{\theta}_i^{\text{IS}}$	$\tilde{\theta}_i^{\text{BLP}}$	δ_i^{M}
θ_1-θ_3	0	0.91	0.91	0.74	0.56
θ_4, θ_5	1	1.27	1.27	1.30	1.22
θ_6, θ_7	2	1.71	1.71	1.85	1.89
θ_8	3	2.21	2.21	2.41	2.46
θ_9	6	4.06	4.06	4.07	4.56

parameters to be estimated. We then assume that each mean θ_i independently follows a LN(μ, σ^2) distribution, or equivalently that the parameters $\gamma_i = \log(\theta_i)$, $i = 1, \ldots, 9$, are independently and identically distributed as normal $N(\mu, \sigma^2)$ random variables.

We estimate the expected number of auditing errors applying the two methods developed in the preceding sections, that is the BLP in (3.13) and the importance sampling estimator (3.18). Here, as the prior mean ξ and the prior variance ϕ, or equivalently the hyperparameters μ and σ^2 are not known, we follow an EB approach for their estimation. Thus, we substitute the parameters ξ, ϕ, μ and σ^2 involved in (3.13) and (3.18) by their method of moments estimates, given in (3.22) and (3.23). Using the data in Table 3.1 we obtain the sample mean $\bar{y} = 1.667$ and the sample variance $s^2 = 3.75$. Hence, we can estimate the prior mean as $\hat{\xi} = \bar{y} = 1.667$ and the prior variance as $\hat{\phi} = s^2 - \bar{y} = 2.083$. Then, solving for the parameters of the log-normal distribution we calculate the estimates $\widehat{\mu} = 0.231$ and $\widehat{\sigma}^2 = 0.559$. The results from the implementation of the estimates are reported in Table 3.2. The importance sampling estimates were obtained with 10^5 Monte Carlo simulations. The usual ML estimates are given by the observed numerical values. For comparison reasons, we also report the estimates obtained using the EB Morris' estimator $\delta_i^M(\mathbf{Y})$ which was reviewed in Chapter 2 and is given in (2.34). Finally, we give the posterior means $\mathrm{E}(\theta_i|\mathbf{y})$, as these were derived with the use of numerical integration from (3.5) and employing the $\widehat{\mu}$, $\widehat{\sigma}^2$ estimates obtained above.

We first notice that the importance sampling estimates are identical to two decimal places to the posterior means, verifying that the method produces an excellent approximation to the exact solution. The BLP in the third column of

Table 3.2 shrinks the MLE towards the sample mean $\bar{y} = 1.667$ with a weight equal to $c = \frac{\hat{\xi}}{\hat{\xi}+\hat{\phi}} = 0.444$, meaning that the BLP estimates adjust the observed values towards the sample mean using a 44.4% proportion of the distance $(y_i - \bar{y})$. As expected, Morris' $\delta_i^M(\mathbf{Y})$ method shrinks less than the BLP, multiplying the smoothing coefficient by a factor equal to $\frac{m-3}{m-1} = 0.75$, to account for the underestimation of the posterior variance occurred with EB estimation.

It is also remarkable that although the linear predictor is constructed in a way such that it adjusts the MLE towards a data estimate of the prior mean, that is the sample mean $\bar{y}$, the shrinking direction of the importance sampling estimator is not evident. The method seems to smooth the observed values towards a central point in the data range, but we cannot definitely determine whether this point is the sample mean or not. We will return to this question with more examples in a subsequent section.

3.7 Example: Oilwell discoveries data

The oilwell discoveries data set was introduced by Clevenson and Zidek (1975), and refers to the number of oilwell discoveries obtained from wildcat exploration in Alberta, Canada, for 36 months during the period 1953–1970 (March and September of each year). The observations are given in Table 3.3. Again, we assume that each observation is an independent Poisson random variable Y_i conditional on its respective mean θ_i, $i = 1, \ldots, 36$, and that the Poisson means independently follow a $\text{LN}(\mu, \sigma^2)$ distribution. The data have also been analysed by Leonard (1976) and George, Makov and Smith (1994).

Table 3.3: *Oilwell discoveries data (Clevenson and Zidek, 1975).*

Oilwell discoveries y	0	1	2	3	5
Observed frequency	19	10	4	2	1

The expected number of discoveries θ_i, $i = 1, \ldots, 36$, is estimated using the linear predictor (3.13) and the importance sampling method (3.18). Following the same EB approach as for the audit data example in the preceding section, we estimate the prior mean ξ and the prior variance ϕ using the data. For this data set the sample mean is $\bar{y} = 0.806$ and the sample variance $s^2 = 1.304$, and therefore we obtain the estimates $\hat{\xi} = \bar{y} = 0.806$ and $\hat{\phi} = s^2 - \hat{\xi} = 0.498$. The parameters of the $\text{LN}(\mu, \sigma^2)$ distribution are then estimated as $\widehat{\mu} = -0.501$ and $\widehat{\sigma}^2 = 0.570$. Substituting these values in (3.13) and (3.18) we obtain the

Table 3.4: *EB estimates of the expected number of discoveries* θ_i, $i = 1, \ldots, 36$, *in the oilwell data example.*

		EB estimates			
par.	y	*post. mean*	$\tilde{\theta}_i^{\text{IS}}$	$\tilde{\theta}_i^{\text{BLP}}$	δ_i^{A}
θ_1-θ_{19}	0	0.55	0.55	0.50	0.31
θ_{20}-θ_{29}	1	0.81	0.81	0.88	0.92
θ_{30}-θ_{33}	2	1.16	1.16	1.26	1.54
θ_{34}, θ_{35}	3	1.58	1.58	1.64	2.15
θ_{36}	5	2.63	2.64	2.41	3.37

results shown in Table 3.4. Again, the importance sampling method estimates, derived with 10^5 Monte Carlo simulations are compared against the EB posterior means calculated with numerical integration. In this data set, the large number of Poisson means ($m = 36$) implies that Morris' method (2.34), which was used in the audit data example, will only differ slightly from the BLP, multiplying the latter's shrinking proportion by a factor equal to $\frac{m-3}{m-1} = 0.94$. Therefore, this method will not be included in the analysis, and instead we present the estimates obtained with Albert type estimator given in (2.36).

The conclusions drawn from the estimates in Table 3.4 are similar to those for the audit data example. The EB importance sampling estimates show that we can rely on this Monte Carlo simulation technique to obtain an accurate and fast approximation to the posterior mean. The BLP shrinks the observed value y_i towards the estimated prior mean $\hat{\xi} = 0.806$ with a weight equal to $c = \frac{\hat{\xi}}{\hat{\xi}+\hat{\phi}} = 0.618$, reflecting a relatively strong belief in the prior specification as this is expressed by the $\hat{\phi} = 0.498$ estimate for the prior variance. On the other hand, the estimates produced with the Albert type method (2.36), reported in the last column of Table 3.4, suggest that this method yields conservative shrinkages compared to both the BLP and the importance sampling estimator, attaching less weight to the prior information, as a result of the largest observation being far from the estimated prior mean.

Again, as with the audit data example, the importance sampling approximation to the posterior mean of the Poisson parameters θ_i, $i = 1, \ldots, 36$, shrinks the MLE towards a central value. In the following section we consider simulated data sets in an effort to investigate whether this central value is the sample mean.

3.8 Shrinking behaviour of the EB posterior mean

In both the audit data and the oilwell discoveries examples, the importance sampling approximation to the posterior mean of the Poisson parameters θ_i, $i = 1, \ldots, m$, appears to adjust the MLE towards a central value. The BLP and a large part of the EB methodology, as given by several authors and reviewed in Chapter 2, suggest that this value should be the sample mean. However, our experimentation with a large number of simulated data sets, showed that unlike the estimators proposed in the literature, the EB posterior mean, as approximated by the importance sampling estimator, shrinks the MLE towards a central point, which lies between the minimum observation and the sample mean, being close to the latter. In the following paragraph we consider two artificial data sets which illustrate this shrinkage behaviour.

Examples: Further data

To obtain the two data sets we first generated m independent random variables θ_i, $i = 1, \ldots, m$, from a log-normal $\text{LN}(\mu, \sigma^2)$ distribution with arbitrarily chosen parameters μ and σ^2. Then, for each θ_i, an observation y_i, $i = 1, \ldots, m$, was simulated independently from a Poisson(θ_i) distribution. The number of the Poisson means is $m = 9$ and $m = 15$ for the first and second data set respectively. The data, together with the EB estimates using the importance sampling estimator and the BLP are given in Tables 3.5 and 3.6.

Table 3.5: *EB estimates of the Poisson parameters θ_i, $i = 1, \ldots, m$, in the first simulated data example. The sample mean is equal to $\bar{y} = 5.222$ and the sample variance is $s^2 = 13.944$.*

		EB estimates	
par.	y	$\tilde{\theta}_i^{\text{IS}}$	$\tilde{\theta}_i^{\text{BLP}}$
θ_1	1	2.90	2.58
θ_2,θ_3	2	3.35	3.21
θ_4	3	3.84	3.83
θ_5,θ_6	5	4.93	5.08
θ_7	8	6.76	6.96
θ_8	9	7.40	7.58
θ_9	12	9.50	9.46

Table 3.6: *EB estimates of the Poisson parameters* θ_i, $i = 1, \ldots, m$, *in the second simulated data example. The sample mean is equal to* $\bar{y} = 9.267$ *and the sample variance is* $s^2 = 24.638$.

		EB estimates	
par.	y	$\tilde{\theta}_i^{\text{IS}}$	$\tilde{\theta}_i^{\text{BLP}}$
θ_1,θ_2	4	6.21	5.98
θ_3	5	6.72	6.60
θ_4,θ_5	6	7.24	7.23
θ_6,θ_7	7	7.79	7.85
θ_8	8	8.33	8.48
θ_9	9	8.93	9.10
θ_{10},θ_{11}	10	9.55	9.72
θ_{12},θ_{13}	12	10.77	10.97
θ_{14}	17	14.14	14.09
θ_{15}	22	17.61	17.21

The sample mean of the first simulated data set, in Table 3.5 is $\bar{y} = 5.222$ and the sample variance is $s^2 = 13.944$. The data estimates for the prior mean and variance are $\hat{\xi} = 5.222$ and $\hat{\phi} = 8.722$ respectively. We notice that the BLP estimate for θ_5 is $\tilde{\theta}_5^{\text{BLP}} = 5.08$, obtained with observation $y_5 = 5$ adjusted towards the estimated prior mean $\hat{\xi} = 5.222$ with a weight equal to $c = \frac{\hat{\phi}}{\hat{\xi}+\hat{\phi}} = 0.375$. However, the importance sampling estimate for the same parameter is equal to $\tilde{\theta}_5^{\text{IS}} = 4.93$, implying that the EB posterior mean shrinks the observed value $y_5 = 5$ towards the opposite direction of that of the sample mean. The estimates for the second data set, reported in Table 3.6 reveal a similar shrinking behaviour. The importance sampling estimate for θ_9 is $\tilde{\theta}_9^{\text{IS}} = 8.93$, despite the fact that the estimated prior mean is $\hat{\xi} = \bar{y} = 9.267$.

3.9 Frequency properties of the EB estimators

The two methods developed in this chapter for the simultaneous estimation of m Poisson means θ_i, $i = 1, \ldots, m$, under a log-normal prior structure, provide approximations to the mean of the posterior distribution of θ_i when an EB solution is adopted for the estimation of the parameters of the log-normal prior distribution. The posterior mean is the Bayes rule under the assumed prior dis-

tribution with respect to the squared error loss function. Therefore, it minimises the average risk, giving excellent frequency properties especially in the region where $\boldsymbol{\theta}$ is more likely to lie, that is in the region to which the prior distribution attaches high probability. Thus, both the BLP and the importance sampling estimator are expected to possess good frequency properties, since the former is an approximation to the posterior mean constructed in such a way that it has the minimum average risk among the linear estimators of its form, and the latter was empirically shown to produce a highly accurate approximation.

Many of the shrinkage methods proposed in the literature for the simultaneous estimation of several Poisson means, smooth the observed values towards a fixed point and therefore provide considerable improvement over the risk of the MLE only when the true unknown parameters are close to the chosen point of shrinkage. However, estimators that exhibit low risk for a variety of different values of the parameter vector $\boldsymbol{\theta}$ are preferred. Thus, we want to assess the frequency properties of the estimators under consideration in a wide range of the parameter space. We examine their average risk when the Poisson parameters θ_i, $i = 1, \ldots, m$, are small, being close to the origin, moderate or relatively large. Also, since the magnitude of shrinkage depends on the variation in the parameters θ_i, $i = 1, \ldots, m$, as suggested by the shrinking coefficient $c = \frac{\mathrm{E}(\theta)}{\mathrm{var}(\theta)+\mathrm{E}(\theta)}$ of the BLP, we are interested in assessing the risk behaviour of the estimators when θ_i, $i = 1, \ldots, m$, are close to each other, as well as when they lie in a wider range.

3.9.1 Loss functions

The posterior mean has minimum average risk under the quadratic loss function. However, the selection of a unique suitable loss function for a specific estimation problem is not always an easy task and hence, estimators that perform well under various loss functions are often desired. We therefore consider different loss functions for the evaluation of the average risk of the examined methods.

We suppose that the parameters θ_i, $i = 1, \ldots, m$ are estimated using δ_i, $i = 1, \ldots, m$. The summed quadratic loss functions

$$\mathrm{SEL}_k = \sum_{i=1}^{m} \frac{(\delta_i - \theta_i)^2}{\theta_i^k}, \tag{3.25}$$

where $k = 0, 1, 2$, which are widely used in the literature and under which most of the methods reviewed in Chapter 2 were developed, are initially considered. Another family of loss functions, often encountered in the literature, is the absolute error loss. It is interesting to examine how our estimators will behave under this loss function, since in this case the posterior median rather than the

posterior mean would be expected to exhibit the lowest average risk with respect to the assumed prior distribution. In accordance to the weighted squared error loss functions in (3.25), we will assess the frequency properties of the considered estimators under the absolute error loss functions

$$\mathrm{AEL}_k = \sum_{i=1}^{m} \frac{|\delta_i - \theta_i|}{\theta_i^k}, \tag{3.26}$$

for $k = 0, \frac{1}{2}$ and 1.

The loss functions considered until now, have the common characteristic of summing over the m components of the vector parameter $\boldsymbol{\theta}$. In essence, the possibility of improving upon the risk of the MLE is based on combining the strength of all the components of the estimator, and in fact the MLE has been shown to be inadmissible, only under loss functions that sum over at least two Poisson means (e.g. Clevenson and Zidek, 1975, Tsui and Press, 1982). We wish to investigate the behaviour of the estimators in the case that the maximum component of a squared error loss function is considered instead of the sum of all components. We therefore employ the maximum component loss functions

$$\mathrm{MAXSEL}_k = \max_{1 \le i \le m} \left\{ \frac{(\delta_i - \theta_i)^2}{\theta_i^k} \right\}, \tag{3.27}$$

for $k = 0, 1, 2$.

The average risk of the nonlinear importance sampling estimator under the considered loss functions can only be computed empirically, employing Monte Carlo simulations. This is also the case for the BLP, except when the squared error loss function $\mathrm{SEL}_0 = \sum_{i=1}^{m} (\delta_i - \theta_i)^2$ is involved. In this case, we can derive an exact analytical expression for the average risk, as follows. We first write the average risk as the expectation of the frequentist risk with respect to the prior distribution of θ, i.e.

$$R_\pi \left(\widetilde{\boldsymbol{\theta}}^{\mathrm{BLP}} \right) = \mathrm{E}_\theta \left\{ R \left(\widetilde{\boldsymbol{\theta}}^{\mathrm{BLP}} \right) \right\}, \tag{3.28}$$

where $R\left(\widetilde{\theta}^{\mathrm{BLP}}\right)$ is given by

$$R \left(\widetilde{\boldsymbol{\theta}}^{\mathrm{BLP}} \right) = \mathrm{E}_{Y|\theta} \left\{ \sum_{i=1}^{m} \left(\tilde{\theta}_i^{\mathrm{BLP}} - \theta_i \right)^2 \right\},$$

and using the form of the BLP in (3.14) we obtain

$$R \left(\widetilde{\boldsymbol{\theta}}^{\mathrm{BLP}} \right) = \sum_{i=1}^{m} \left[\mathrm{E}_{Y|\theta} \{ (1-c) Y_i + c\xi - \theta_i \}^2 \right]$$

which gives

$$R\left(\widetilde{\boldsymbol{\theta}}^{\mathrm{BLP}}\right) = \sum_{i=1}^{m}\left[\{c(\xi-\theta_i)\}^2 + (1-c)^2\theta_i\}\right]$$
$$= c^2\sum_{i=1}^{m}(\theta_i-\xi)^2 + (1-c)^2 m\bar{\theta},$$

where $\bar{\theta} = m^{-1}\sum_{i=1}^{m}\theta_i$. Then, substituting the frequentist risk from the above expression into (3.28), and using the form of c in (3.15) we obtain

$$R_\pi\left(\widetilde{\boldsymbol{\theta}}^{\mathrm{BLP}}\right) = c^2\mathrm{E}_\theta\left\{\sum_{i=1}^{m}(\theta_i-\xi)^2\right\} + (1-c)^2 m\mathrm{E}_\theta(\bar{\theta})$$
$$= m\left\{c^2\phi + (1-c)^2\xi\right\}$$
$$= \frac{\mathrm{var}(\theta_i)}{\mathrm{var}(\theta_i)+\mathrm{E}(\theta_i)}\, m\,\mathrm{E}(\theta_i). \tag{3.29}$$

We notice here that, as mentioned in Subsection 2.3.1, when the EB estimator (3.24) is considered, the risk of the BLP as $m \to \infty$ is approximately given by $\left\{\frac{S(\boldsymbol{\theta})}{S(\boldsymbol{\theta})+\bar{\theta}}\right\} m\bar{\theta}$, where $\bar{\theta} = \sum_{i=1}^{m}\theta_i/m$ and $S(\boldsymbol{\theta}) = \sum_{i=1}^{m}(\theta_i-\bar{\theta})^2/(m-1)$. The analogy between this expression and (3.29) suggests that the risk function of the considered estimators should be expected to exhibit a similar behaviour to the average risk results obtained in this thesis.

Finally, the average risk of the MLE, to which the considered estimators will be compared, can be analytically obtained under the squared error loss functions SEL_0 and SEL_1. Using he former we obtain

$$R_\pi(\mathbf{Y}) = \mathrm{E}_\theta\mathrm{E}_{Y|\theta}\left\{\sum_{i=1}^{m}(Y_i-\theta_i)^2\right\}$$
$$= \mathrm{E}_\theta\left\{\sum_{i=1}^{m}\mathrm{var}_{Y|\theta}(Y_i)\right\} = m\,\mathrm{E}(\theta_i). \tag{3.30}$$

Under SEL_1 the average risk of the MLE is given by

$$R_\pi(\mathbf{Y}) = \mathrm{E}_\theta\mathrm{E}_{Y|\theta}\left\{\sum_{i=1}^{m}\frac{(Y_i-\theta_i)^2}{\theta_i}\right\}$$
$$= \mathrm{E}_\theta\left[\sum_{i=1}^{m}\left\{\frac{\mathrm{var}_{Y|\theta}(Y_i)}{\theta_i}\right\}\right] = m. \tag{3.31}$$

3.9.2 Frequency simulations

The considered estimators will be assessed and compared using the frequency criterion of average risk, which for an estimator $\boldsymbol{\delta}$ of $\boldsymbol{\theta}$ is given by

$$R_\pi(\boldsymbol{\delta}) = \mathrm{E}_\theta\mathrm{E}_{Y|\theta}\left\{L(\boldsymbol{\delta},\boldsymbol{\theta})\right\},$$

with $L(\boldsymbol{\delta}, \boldsymbol{\theta})$ being one of the loss functions (3.25), (3.26) and (3.27) presented in the preceding subsection. The average risk of each examined estimator will be compared to the corresponding risk of the usual estimator $\mathbf{Y}$, which is both the MLE and the UMVUE. As described in Subsection 3.9.1, the average risk of the MLE and the BLP can be analytically computed under a limited number of specific loss functions. However, for comparison reasons, we will use Monte Carlo simulations in all cases to estimate the average risk. For the simulation procedure, we first chose the true mean ξ and the true variance ϕ of the Poisson parameters, and calculated the corresponding μ and σ^2 log-normal hyperparameters. We then generated a number m of θ_i values, independently from a $\mathrm{LN}(\mu, \sigma^2)$ distribution. For each θ_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, $i = 1, \ldots, m$, were computed according to the estimating methods of interest, together with the associated sum and maximum component of the error loss. With the θ_i, $i = 1, \ldots, m$, fixed, the last step was repeated N_k times for the evaluation of the frequentist risk $R(\boldsymbol{\delta}) = \mathrm{E}_{Y|\theta}\{L(\boldsymbol{\delta}, \boldsymbol{\theta})\}$, and then new θ_i, $i = 1, \ldots, m$, values were sampled and the whole procedure was repeated N_t times to allow the computation of the average risk $R_\pi(\boldsymbol{\delta}) = \mathrm{E}_\theta\{R(\boldsymbol{\delta})\}$. The latter was estimated using

$$\widehat{R}_\pi(\boldsymbol{\delta}) = \frac{1}{N_t N_k} \sum_{t=1}^{N_t} \sum_{k=1}^{N_k} \{L(\boldsymbol{\delta}_{tk}, \boldsymbol{\theta}_t)\}. \tag{3.32}$$

For example, under the $\mathrm{SEL}_0 = \sum_{i=1}^{m} (\delta_i - \theta_i)^2$ loss function, the estimated average risk of an estimator $\boldsymbol{\delta}$ is given by

$$\widehat{R}_\pi(\boldsymbol{\delta}) = \frac{1}{N_t N_k} \sum_{t=1}^{N_t} \sum_{k=1}^{N_k} \left\{ \sum_{i=1}^{m} (\delta_{itk} - \theta_{it})^2 \right\}.$$

To compare the performance of the methods against that of the MLE, we calculated the relative average risk improvement (RARI) of each estimator with respect to MLE. This is defined as

$$\mathrm{RARI}(\boldsymbol{\delta}) = \frac{\widehat{R}_\pi(\mathbf{Y}) - \widehat{R}_\pi(\boldsymbol{\delta})}{\widehat{R}_\pi(\mathbf{Y})} \times 100\%, \tag{3.33}$$

and indicates the percentage of the improvement of the estimator $\boldsymbol{\delta}$ when compared to the MLE in terms of average risk.

In an attempt to investigate the performance of the estimators within a wide range of the $\boldsymbol{\theta}$ parameter space, the simulation procedure was applied to 9 different settings for the Poisson parameters, corresponding to 9 combinations of their mean ξ and their variance ϕ. These, together with the associated μ and σ^2 log-normal parameters, are given in Table 3.7. We notice that the variance is set to

Table 3.7: *Chosen values for the true mean $\xi = E(\theta_i)$, true variance $\phi = var(\theta_i)$ and corresponding hyperparameters μ and σ^2 for the 9 specified combinations.*

moments		*parameters*	
ξ	ϕ	μ	σ^2
1.0	0.5	−0.20	0.41
	1.0	−0.35	0.69
	2.0	−0.55	1.10
5.0	2.5	1.56	0.10
	5.0	1.52	0.18
	10.0	1.44	0.34
10.0	5.0	2.28	0.05
	10.0	2.26	0.10
	20.0	2.21	0.18

be half, equal and twice the magnitude of the mean. In terms of the marginal distribution of the data, this corresponds to the variance of each simulated data set ranging from one and a half times to three times the marginal mean.

A total number of 4×10^4 Monte Carlo simulations ($N_t \times N_k = 200 \times 200$) were used for the estimation of the average risk with (3.32). The simulation study was performed using the C computer programming language. A summarised outline of the algorithm used is shown in Figure 3.1.

3.9.3 Results

The average risk improvement (3.33) of the importance sampling estimator in (3.18) and the BLP in (3.24) when compared to the MLE, for the simultaneous estimation of $m = 10$ Poisson means is given in Tables 3.9 through 3.17. In each table we report the RARI of the considered estimators, corresponding to a specific loss function and calculated using the estimated average risk (3.32), for the 9 combinations of the true moments of θ.

A general conclusion that can be drawn from all tables is that in most of the examined cases the importance sampling estimator and the BLP produce remarkably high relative improvement in average risk in comparison to the MLE. As shown in Table 3.8, the improvement yielded when the importance sampling estimator is used, is over 60% in 12 of the 81 examined cases, and between

```
Choose a pair of ξ, φ values
    for t=1 to N_t (no. of θ samples)
        for i=1 to m
            Generate θ_i ~iid~ LN(μ, σ²)
            for k=1 to N_k (no. of data samples)
                Generate Y_ik ~ Poisson(θ_i)
                Compute estimators
            endfor k
        endfor i
        Calculate risk for estimators
    endfor t
    Calculate average risk and RARI
Repeat with different ξ, φ values
```

Figure 3.1: *Pseudo-code for C program used for the computer simulation study.*

$40\% - 60\%$ in 19 cases. The method fails to improve the MLE in only 6 out of the 81 examined cases. Table 3.8 also demonstrates that the number of times that the BLP improves the MLE with a percentage of $20\% - 60\%$ is similar to that for the importance sampling estimator. However, the latter has at least 60% lower average risk than the MLE 12 times, while the BLP only 7. Overall, the importance sampling estimator performs better than the BLP, producing higher savings than the linear method in 60 out of the 81 considered cases.

Table 3.8: *Summary of the relative improvement in average risk (RARI) of the importance sampling estimator and the BLP when compared to the MLE. A total number of* 81 *cases were examined.*

% *RARI*	< 0	$0 - 20$	$20 - 40$	$40 - 60$	> 60
$\boldsymbol{\theta}^{\mathrm{IS}}$	6	7	33	19	12
$\boldsymbol{\theta}^{\mathrm{BLP}}$	5	11	35	19	7

For comparison reasons we also present the RARI of the methods reviewed in Chapter 2. These are arranged in categories according to the direction of shrinking. Morris' $\boldsymbol{\delta}^{\mathrm{M}}(\mathbf{Y})$ in (2.34) and the Albert-type (2.36) $\boldsymbol{\delta}^{\mathrm{A}}(\mathbf{Y})$ rules smooth the MLE towards the sample mean; Hudson's (2.24) method $\boldsymbol{\delta}^{\mathrm{H}}(\mathbf{Y})$ and the Ghosh, Hwang and Tsui estimators $\boldsymbol{\delta}^{\mathrm{GHT4}}(\mathbf{Y})$ in (2.23) and $\boldsymbol{\delta}^{\mathrm{GHT3}}(\mathbf{Y})$ in (2.22) adjust the MLE towards a central data point, that being approximately the geometric mean for the first two and the median for the third; Tsui's (2.19) estimator $\boldsymbol{\delta}^{\mathrm{T}}(\mathbf{Y})$

and the Ghosh, Hwang and Tsui rules $\boldsymbol{\delta}^{\mathrm{GHT1}}(\mathbf{Y})$ and $\boldsymbol{\delta}^{\mathrm{GHT2}}(\mathbf{Y})$, given in (2.20) and (2.21) respectively, shrink to the minimum observation, with the first two developed under the SEL_0 loss function and the third under SEL_1; finally, Peng's $\boldsymbol{\delta}^{\mathrm{P}}(\mathbf{Y})$ estimator in (2.18), the Clevenson and Zidek (2.25) method $\boldsymbol{\delta}^{\mathrm{CZ}}(\mathbf{Y})$ and the Tsui and Press $\boldsymbol{\delta}^{\mathrm{TP}}(\mathbf{Y})$ estimator in (2.26) smooth the usual estimate towards zero and are constructed under SEL_0, SEL_1 and SEL_2 respectively.

When the $\mathrm{SEL}_0 = \sum_{i=1}^{m}(\delta_i - \theta_i)^2$ loss function is used, both the importance sampling estimator and the BLP perform outstandingly. The results in Table 3.9 reveal that the importance sampling method produces greater improvement than

Table 3.9: *Percentage of relative improvement in average risk (3.33) under* SEL_0 *when the estimators are compared to the MLE.*

$\mathrm{E}(\theta_i)$	1.0			5.0			10.0		
$\mathrm{var}(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{\mathrm{IS}}$	62.9	51.5	37.8	60.6	46.1	26.7	61.1	42.7	26.7
$\boldsymbol{\theta}^{\mathrm{BLP}}$	60.7	49.3	36.0	59.9	45.3	25.8	60.7	42.1	26.1
Shrink to sample mean									
$\boldsymbol{\delta}^{\mathrm{M}}$	56.1	46.7	35.4	55.4	43.3	26.4	56.0	40.7	26.7
$\boldsymbol{\delta}^{\mathrm{A}}$	52.8	45.6	36.2	52.3	43.0	28.2	52.7	41.0	28.6
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{\mathrm{H}}$	35.4	27.4	19.6	50.1	37.5	24.1	52.5	39.4	25.5
$\boldsymbol{\delta}^{\mathrm{GHT4}}$	41.2	34.2	26.5	31.4	24.9	18.0	31.7	25.0	17.2
$\boldsymbol{\delta}^{\mathrm{GHT3}}$	15.3	14.0	12.4	7.9	7.4	6.3	3.7	3.6	3.3
Shrink to minimum observation									
$\boldsymbol{\delta}^{\mathrm{GHT1}}$	22.1	19.7	16.5	12.7	10.2	7.3	13.4	10.5	7.1
$\boldsymbol{\delta}^{\mathrm{T}}$	25.9	22.7	18.8	6.4	5.6	4.5	3.6	3.2	2.5
$\boldsymbol{\delta}^{\mathrm{GHT2}}$	61.4	52.5	41.9	28.5	24.3	18.9	20.5	17.7	13.7
Shrink to zero									
$\boldsymbol{\delta}^{\mathrm{P}}$	25.2	22.3	18.5	3.9	3.6	3.2	1.2	1.1	1.0
$\boldsymbol{\delta}^{\mathrm{CZ}}$	45.2	37.1	26.6	16.9	14.2	10.9	9.5	8.3	6.7
$\boldsymbol{\delta}^{\mathrm{TP}}$	22.9	−25.7	−91.0	33.9	19.0	1.2	25.2	19.8	8.9

all the other EB estimators, except in the cases when $\mathrm{E}(\theta_i) = 5$, $\mathrm{var}(\theta_i) = 10$ and $\mathrm{E}(\theta_i) = 10$, $\mathrm{var}(\theta_i) = 20$, where the Albert-type estimator performs better. This is not surprising, since, as described in Subsection 2.3.1, the latter is derived in a way such that it behaves better than the posterior mean when the observations are not close to the prior mean. The BLP also seems to dominate the Morris and Albert-type estimators when the variation in θ_i is small to moderate. However, when the θ_i, $i = 1, \ldots, m$, are not close to each other, $\boldsymbol{\delta}^{\mathrm{M}}(\mathbf{Y})$ and $\boldsymbol{\delta}^{\mathrm{A}}(\mathbf{Y})$ restrict the shrinking towards the prior mean, which can give better average risk properties when overdispersed observations occur. In general, the RARI for all estimators drops as the θ_i, $i = 1, \ldots, m$, are less concentrated around their mean. Large variation in the true parameters leads to higher estimates for the prior variance, which in turn implies small confidence in the prior structure. The latter is then reflected in less shrinking towards the prior mean. Thus, as $\mathrm{var}(\theta_i)$ gets larger, the EB methods give estimates which are closer to the MLE, producing average risk that tends to be similar to that of the usual estimator.

The performance of the remaining linear estimators is much less impressive than that of the EB rules, with only a few exceptions. Hudson's estimator $\boldsymbol{\delta}^{\mathrm{H}}(\mathbf{Y})$, which approximately shrinks towards the geometric mean of the data does well, especially when the Poisson parameters θ_i, $i = 1, \ldots, m$, are not close to the origin. On the other hand, the Ghosh, Hwang and Tsui $\boldsymbol{\delta}^{\mathrm{GHT4}}(\mathbf{Y})$ method, which also approximately smoothes towards the geometric mean, performs better for small θ_i. Estimator $\boldsymbol{\delta}^{\mathrm{GHT3}}(\mathbf{Y})$ does not give substantial savings in average risk, despite the fact that it also shrinks to a central data value, namely the median of the data. From the methods adjusting the observed value towards the minimum observation or zero, estimators $\boldsymbol{\delta}^{\mathrm{GHT2}}(\mathbf{Y})$ and $\boldsymbol{\delta}^{\mathrm{CZ}}(\mathbf{Y})$ produce good average risk results, but only when the θ_i, $i = 1, \ldots, m$, are close to the origin, as expected from the fact that they shrink towards small observation values.

Table 3.10 contains the RARI results when the $\mathrm{SEL}_1 = \sum_{i=1}^{m} \frac{(\delta_i - \theta_i)^2}{\theta_i}$ loss function is involved. Again, as with SEL_0, the nonlinear importance sampling estimator dominates the other methods in most of the various $\boldsymbol{\theta}$ regions. The linear EB estimators $\boldsymbol{\delta}^{\mathrm{M}}(\mathbf{Y})$ and $\boldsymbol{\delta}^{\mathrm{A}}(\mathbf{Y})$ have better average risk properties than both the importance sampling approximation to the posterior mean and the BLP when the Poisson parameters θ_i, $i = 1, \ldots, m$, are widely scattered around their mean, due again to their reduced shrinking in overdispersed situations. The average risk behaviour of all the examined estimators follows a similar comparative pattern to that under the SEL_0 loss function. However, while the average risk relative savings of the importance sampling estimator and the BLP under SEL_0 and SEL_1 are similar for θ_i, $i = 1, \ldots, m$, not being close to zero, there seems

Table 3.10: *Percentage of relative improvement in average risk (3.33) under* SEL_1 *when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	60.3	45.5	15.5	60.9	46.0	25.0	61.1	42.8	26.7
$\boldsymbol{\theta}^{BLP}$	58.1	44.7	21.3	59.8	44.8	24.4	60.6	41.9	25.9
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	55.6	45.9	29.8	55.6	43.7	26.9	56.1	41.0	27.3
$\boldsymbol{\delta}^{A}$	52.3	45.1	32.2	52.4	43.3	28.7	52.7	41.2	29.2
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	35.3	26.3	15.4	50.4	37.9	19.2	52.7	39.6	25.0
$\boldsymbol{\delta}^{GHT4}$	44.2	37.1	27.0	33.2	28.0	21.7	32.7	26.7	19.4
$\boldsymbol{\delta}^{GHT3}$	17.0	17.0	17.0	8.5	8.5	8.2	3.9	3.9	3.9
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	25.1	25.2	24.8	13.0	11.5	9.8	13.5	11.3	8.8
$\boldsymbol{\delta}^{T}$	29.0	28.7	28.0	6.8	6.8	6.6	3.8	3.6	3.5
$\boldsymbol{\delta}^{GHT2}$	60.5	49.5	29.5	28.8	26.4	23.4	20.4	18.2	15.5
Shrink to zero									
$\boldsymbol{\delta}^{P}$	29.1	29.3	29.0	3.8	4.1	4.7	1.0	1.0	1.2
$\boldsymbol{\delta}^{CZ}$	51.9	52.4	53.1	16.9	16.0	15.3	9.1	8.5	8.2
$\boldsymbol{\delta}^{TP}$	39.7	28.2	14.9	35.3	27.8	19.7	25.2	21.6	16.2

to be a decrease when $E(\theta_i) = 1$ and SEL_1 is involved. This is more evident when the variance $var(\theta_i)$ is large, and can be explained from the fact that the $SEL_1 = \sum_{i=1}^{m} \frac{(\delta_i - \theta_i)^2}{\theta_i}$ loss function penalises heavily the cases where θ_i is very close to zero. It is also interesting to examine the performance of the two estimators constructed to universally dominate the MLE under the considered SEL_1 loss function, that is the Ghosh, Hwang and Tsui $\boldsymbol{\delta}^{GHT2}(\mathbf{Y})$ and the Clevenson and Zidek $\boldsymbol{\delta}^{CZ}(\mathbf{Y})$ estimators. They both produce very large RARI when $E(\theta_i) = 1$, with the Clevenson and Zidek method dominating all estimators when $var(\theta_i) = 1$ or 2. However, since they shrink towards the minimum observation and zero, their performance drops dramatically as $E(\theta_i)$ increases, and therefore the importance sampling and linear methods developed in this chapter offer better average risk properties overall.

When the $SEL_2 = \sum_{i=1}^{m} \frac{(\delta_i - \theta_i)^2}{\theta_i^2}$ loss function is considered, we expect a considerable effect caused by small θ_i values attaching more weight to large discrepancies between δ_i and θ_i. The RARI results in Table 3.11 show that the importance sampling estimator does not perform well when the variance of θ_i is large, and can be much worse than the MLE when θ_i, $i = 1, \ldots, m$, are very close to zero. The linear EB estimators compare better to the MLE when the true variance $var(\theta_i)$ is moderate to large. It is also remarkable that the rules shrinking towards zero or $y_{(1)}$ perform very well when the θ_i, $i = 1, \ldots, m$, are close to the origin and have moderate to large variance. We notice that when $E(\theta_i) = 1$, the Clevenson and Zidek $\boldsymbol{\delta}^{CZ}(\mathbf{Y})$ estimator gives greater RARI with SEL_2 than under SEL_1, despite the fact that it is developed to dominate universally the MLE under the latter. The Tsui and Press $\boldsymbol{\delta}^{TP}(\mathbf{Y})$ estimator, which dominates the MLE under SEL_2, in addition to the advantage of never being worse than the usual estimate, it also gives greater RARI than the importance sampling estimator when the variation of θ_i is large. However, it is dominated by the the importance sampling method and the other EB rules under SEL_2 in more of the examined cases.

We will now examine the EB performance of the two approximations to the posterior mean of θ_i, that is the importance sampling method and the BLP, under the absolute error loss functions (3.26). The RARI results reported in Tables 3.12, 3.13 and 3.14 show the same general patterns in the comparison of the examined estimators, as when the SEL_k loss functions in (3.25) were considered. The nonlinear importance sampling approximation to the posterior mean produces greater average risk savings than most of the other methods under consideration, with the BLP following in performance. As the variance in the Poisson parameters increases, the RARI of all estimators drops, with the Morris $\boldsymbol{\delta}^{M}(\mathbf{Y})$ and the Albert-type $\boldsymbol{\delta}^{A}(\mathbf{Y})$ EB rules performing better. The weighted loss functions

Table 3.11: *Percentage of relative improvement in average risk (3.33) under SEL_2 when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	42.3	−8.5	−175.5	58.8	37.3	1.4	60.0	38.7	17.5
$\boldsymbol{\theta}^{BLP}$	42.0	0.2	−119.4	57.4	36.5	5.6	59.2	37.7	17.6
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	46.5	19.9	−54.4	54.5	39.1	16.0	55.4	38.6	22.6
$\boldsymbol{\delta}^{A}$	46.4	27.0	−31.1	51.6	39.9	18.9	52.3	39.5	24.9
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	29.7	9.8	−35.6	48.8	31.6	−7.8	52.0	36.7	17.3
$\boldsymbol{\delta}^{GHT4}$	41.0	22.6	−17.7	34.6	29.7	21.7	33.5	27.6	19.9
$\boldsymbol{\delta}^{GHT3}$	18.4	18.5	17.4	9.3	9.7	10.1	4.1	4.3	4.6
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	28.2	30.1	31.6	13.3	12.6	11.9	13.6	12.0	10.2
$\boldsymbol{\delta}^{T}$	32.0	33.7	35.2	7.3	8.0	8.9	4.0	4.1	4.5
$\boldsymbol{\delta}^{GHT2}$	51.5	18.3	−60.2	29.0	27.6	25.8	20.2	18.4	16.3
Shrink to zero									
$\boldsymbol{\delta}^{P}$	33.0	35.6	37.5	3.8	4.7	6.7	0.8	0.9	1.5
$\boldsymbol{\delta}^{CZ}$	56.8	60.8	65.2	16.8	17.3	18.9	8.7	8.6	9.4
$\boldsymbol{\delta}^{TP}$	40.5	32.3	23.3	35.6	31.2	26.2	24.9	22.5	19.6

Table 3.12: *Percentage of relative improvement in average risk (3.33) under* AEL_0 *when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	42.5	34.6	24.8	38.7	28.5	16.3	38.8	25.2	15.5
$\boldsymbol{\theta}^{BLP}$	40.6	33.2	24.5	37.9	27.6	15.6	38.5	24.6	15.0
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	37.8	32.1	24.9	34.7	26.3	16.0	35.0	23.9	15.4
$\boldsymbol{\delta}^{A}$	35.0	30.8	25.1	32.1	25.8	17.0	32.2	24.0	16.4
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	23.4	18.4	13.5	31.5	22.9	14.1	32.5	23.3	14.8
$\boldsymbol{\delta}^{GHT4}$	29.4	25.0	19.8	19.1	15.4	11.5	18.8	14.7	10.2
$\boldsymbol{\delta}^{GHT3}$	8.9	8.3	7.6	4.3	4.1	3.8	1.9	1.9	1.8
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	11.0	10.5	9.5	5.4	4.4	3.3	6.1	4.8	3.3
$\boldsymbol{\delta}^{T}$	12.9	12.0	10.7	2.5	2.3	2.0	1.4	1.2	1.0
$\boldsymbol{\delta}^{GHT2}$	41.5	35.4	27.2	14.5	12.4	9.9	10.3	8.8	6.8
Shrink to zero									
$\boldsymbol{\delta}^{P}$	12.5	12.1	10.9	0.7	0.7	0.7	0.1	0.0	0.0
$\boldsymbol{\delta}^{CZ}$	23.0	20.9	18.0	6.4	5.3	4.2	3.4	2.9	2.3
$\boldsymbol{\delta}^{TP}$	16.2	5.5	−6.7	17.9	10.8	3.7	12.5	9.6	5.1

Table 3.13: *Percentage of relative improvement in average risk (3.33) under* $AEL_{\frac{1}{2}}$ *when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	41.2	31.2	14.2	38.7	28.2	15.4	38.8	25.1	15.3
$\boldsymbol{\theta}^{BLP}$	39.6	30.9	17.5	37.8	27.3	14.9	38.4	24.4	14.7
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	38.1	31.8	21.7	34.8	26.4	16.1	35.0	23.9	15.5
$\boldsymbol{\delta}^{A}$	35.8	31.3	23.1	32.3	26.0	17.1	32.2	24.0	16.6
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	23.9	18.5	12.5	31.6	22.9	12.7	32.6	23.4	14.6
$\boldsymbol{\delta}^{GHT4}$	31.2	26.3	19.4	19.8	16.5	12.9	19.2	15.3	10.9
$\boldsymbol{\delta}^{GHT3}$	9.6	9.3	8.6	4.5	4.5	4.5	2.0	2.0	2.0
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	12.0	11.9	11.2	5.5	4.8	3.9	6.2	5.0	3.8
$\boldsymbol{\delta}^{T}$	13.9	13.5	12.7	2.6	2.7	2.6	1.4	1.4	1.3
$\boldsymbol{\delta}^{GHT2}$	41.9	33.7	19.7	14.6	13.1	11.2	10.3	8.9	7.3
Shrink to zero									
$\boldsymbol{\delta}^{P}$	13.9	13.9	13.2	0.6	0.7	1.0	0.0	0.0	0.1
$\boldsymbol{\delta}^{CZ}$	25.2	25.1	24.4	6.3	5.7	5.3	3.2	2.9	2.7
$\boldsymbol{\delta}^{TP}$	18.1	12.1	6.1	18.3	13.2	8.2	12.5	10.2	6.9

Table 3.14: *Percentage of relative improvement in average risk (3.33) under* AEL_1 *when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	36.9	20.3	−18.4	38.3	26.6	11.3	38.6	24.3	13.8
$\boldsymbol{\theta}^{BLP}$	36.1	22.3	−7.2	37.3	25.7	11.8	38.0	23.6	13.4
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	36.8	27.2	5.6	34.7	25.7	14.6	34.9	23.5	14.8
$\boldsymbol{\delta}^{A}$	35.8	28.8	10.3	32.2	25.6	15.8	32.2	23.8	16.0
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	23.6	16.2	6.0	31.4	21.9	8.8	32.6	23.0	13.5
$\boldsymbol{\delta}^{GHT4}$	31.9	24.4	11.3	20.6	17.5	13.8	19.5	15.8	11.4
$\boldsymbol{\delta}^{GHT3}$	10.4	10.0	8.7	4.7	4.9	5.2	2.0	2.1	2.2
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	12.9	12.8	12.1	5.6	5.1	4.6	6.2	5.3	4.2
$\boldsymbol{\delta}^{T}$	14.7	14.5	13.8	2.8	3.0	3.2	1.5	1.5	1.6
$\boldsymbol{\delta}^{GHT2}$	40.5	26.5	−1.9	14.7	13.6	12.3	10.2	9.1	7.7
Shrink to zero									
$\boldsymbol{\delta}^{P}$	15.1	15.3	14.6	0.5	0.7	1.4	−0.1	−0.1	0.1
$\boldsymbol{\delta}^{CZ}$	26.9	27.6	27.8	6.2	6.0	6.3	3.1	2.9	3.0
$\boldsymbol{\delta}^{TP}$	17.8	13.3	9.4	18.6	14.7	11.0	12.4	10.6	8.2

$\text{AEL}_{\frac{1}{2}}$ and AEL_1 penalise heavily bad estimates in the region close to zero, and this is again shown for the importance sampling estimator and the BLP in Table 3.14 for $\text{E}(\theta_i) = 1$, $\text{var}(\theta_i) = 2$. As with the squared error loss functions, Hudson's method $\boldsymbol{\delta}^{\text{H}}(\mathbf{Y})$ performs well for large θ_i, $i = 1, \ldots, m$, while the Ghosh, Hwang and Tsui rules $\boldsymbol{\delta}^{\text{GHT4}}(\mathbf{Y})$, $\boldsymbol{\delta}^{\text{GHT2}}(\mathbf{Y})$ and the Clevenson and Zidek estimator $\boldsymbol{\delta}^{\text{CZ}}(\mathbf{Y})$ give good results when $\text{E}(\theta_i)$ is close to zero. We finally notice that all methods yield smaller relative savings in average risk than they did under the squared error loss functions (3.25). This is due to the fact that with the absolute error loss, bad estimates pay a lower loss price than with squared error loss, and therefore the MLE seems to benefit more when the former loss function is preferred. Also, the decrease in RARI can be explained from the fact that nearly all estimators are developed with the aim to offer good risk properties specifically under a squared error loss function.

Until now we have examined the average risk of the considered methods under summed loss functions, meaning that in the loss evaluation we sum over all the m components of the difference $(\delta_i - \theta_i)$, $i = 1, \ldots, m$. It is interesting to investigate the average risk behaviour of the estimators under component-wise loss functions. We consider the worst case for each estimator, by evaluating the average risk when the maximum discrepancy between δ_i and θ_i occurs, for $i = 1, \ldots, m$. In Tables 3.15 through 3.17 we present the RARI results for the considered methods under the loss functions $\text{MAXSEL}_k = \max_{1 \le i \le m} \left\{ \frac{(\delta_i - \theta_i)^2}{\theta_i^k} \right\}$, $k = 0, 1, 2$, respectively.

The general impression given from these tables is that the considered estimators, and especially the importance sampling estimator and the BLP which are the methods of main interest, still produce remarkably high average risk improvement when compared to the MLE. The relative savings can be up to 63% for the importance sampling estimator, and the comparison among the examined methods gives, in most cases, the same conclusions as with the summed squared error loss functions. However, it is interesting to notice that the results shown in Table 3.15 point out that under the MAXSEL_0 loss function the BLP gives larger RARI than the importance sampling estimator in 5 out of the 9 settings for $\boldsymbol{\theta}$ that we consider. The reason for that seems to be that the maximum component of the loss function is likely to occur when an outlier observation Y_i is involved, in which case the importance sampling estimator can shrink less than the BLP and therefore result in larger component-wise average risk. It is also important to notice the excellent performance of the Ghosh, Hwang and Tsui estimator $\boldsymbol{\delta}^{\text{GHT2}}(\mathbf{Y})$ and the Clevenson and Zidek method $\boldsymbol{\delta}^{\text{CZ}}(\mathbf{Y})$ under the MAXSEL_1 and MAXSEL_2 loss functions and when $\text{E}(\theta_i) = 1$. Also, the Tsui and Press $\boldsymbol{\delta}^{\text{TP}}(\mathbf{Y})$ rule dominates the EB estimators under MAXSEL_2 in most of the settings for the true

Table 3.15: *Percentage of relative improvement in average risk (3.33) under $MAXSEL_0$ when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	62.2	49.0	33.7	59.5	43.1	21.9	60.8	40.8	23.9
$\boldsymbol{\theta}^{BLP}$	60.8	47.7	32.6	59.7	43.2	21.8	61.0	40.9	24.1
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	55.6	44.8	32.2	55.1	41.5	23.0	56.2	39.7	24.9
$\boldsymbol{\delta}^{A}$	53.0	44.5	34.1	52.2	41.8	25.3	52.9	40.2	27.0
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	32.9	25.0	17.2	48.1	34.5	22.5	51.7	37.8	23.1
$\boldsymbol{\delta}^{GHT4}$	36.8	29.7	22.3	28.6	21.3	14.7	30.1	22.9	14.8
$\boldsymbol{\delta}^{GHT3}$	15.0	13.7	11.7	7.4	6.6	5.3	3.7	3.4	2.9
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	25.2	20.8	15.9	17.5	13.2	8.9	17.4	13.2	8.3
$\boldsymbol{\delta}^{T}$	30.1	24.6	18.7	8.9	7.5	5.5	5.3	4.4	3.3
$\boldsymbol{\delta}^{GHT2}$	61.4	50.8	39.3	33.2	27.2	19.9	23.9	20.5	15.2
Shrink to zero									
$\boldsymbol{\delta}^{P}$	29.1	23.6	17.8	8.2	7.0	5.4	3.1	2.8	2.4
$\boldsymbol{\delta}^{CZ}$	52.8	38.7	22.1	25.4	20.6	14.6	15.1	13.3	10.1
$\boldsymbol{\delta}^{TP}$	10.4	−72.7	−176.4	38.7	17.2	−8.0	29.8	23.5	6.8

Table 3.16: *Percentage of relative improvement in average risk (3.33) under $MAXSEL_1$ when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	62.7	49.0	20.7	60.4	43.9	22.3	61.0	41.9	25.5
$\boldsymbol{\theta}^{BLP}$	60.2	47.1	24.6	59.9	43.2	21.6	60.9	41.3	24.9
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	55.8	46.5	31.9	55.4	42.3	24.2	56.3	40.4	26.2
$\boldsymbol{\delta}^{A}$	51.4	44.5	33.6	51.9	42.0	25.9	52.8	40.6	27.9
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	32.6	24.2	13.3	48.3	35.5	12.4	51.8	37.6	21.9
$\boldsymbol{\delta}^{GHT4}$	39.6	34.1	26.2	30.3	24.8	18.2	31.3	25.0	17.7
$\boldsymbol{\delta}^{GHT3}$	17.9	18.6	19.3	8.1	7.8	6.9	4.0	3.8	3.7
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	30.6	29.8	28.9	18.9	16.8	14.5	18.0	15.2	12.2
$\boldsymbol{\delta}^{T}$	36.3	34.9	33.2	10.0	9.9	9.8	5.7	5.4	5.2
$\boldsymbol{\delta}^{GHT2}$	59.0	48.5	32.0	34.2	31.8	28.5	23.7	21.5	18.8
Shrink to zero									
$\boldsymbol{\delta}^{P}$	36.1	35.1	34.1	9.0	9.3	9.6	3.0	3.0	3.3
$\boldsymbol{\delta}^{CZ}$	68.6	68.0	67.4	26.8	25.9	25.1	14.8	14.2	14.0
$\boldsymbol{\delta}^{TP}$	52.5	35.9	16.3	41.2	35.4	28.4	29.7	27.0	22.3

Table 3.17: *Percentage of relative improvement in average risk (3.33) under $MAXSEL_2$ when the considered estimators are compared to the MLE.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
$\boldsymbol{\theta}^{IS}$	33.5	−31.9	−214.4	55.4	23.3	−28.3	58.0	31.7	3.4
$\boldsymbol{\theta}^{BLP}$	33.0	−19.8	−149.2	54.4	23.6	−19.4	57.5	30.9	5.0
Shrink to sample mean									
$\boldsymbol{\delta}^{M}$	39.1	6.5	−72.3	52.1	30.4	−1.3	54.3	34.1	14.0
$\boldsymbol{\delta}^{A}$	38.7	15.9	−44.1	49.3	32.6	2.9	51.3	35.7	16.8
Shrink to geometric mean or median									
$\boldsymbol{\delta}^{H}$	22.1	0.3	−52.2	43.7	20.8	−47.7	49.5	29.6	3.0
$\boldsymbol{\delta}^{GHT4}$	32.0	13.0	−28.5	31.8	25.8	14.2	32.2	25.6	16.8
$\boldsymbol{\delta}^{GHT3}$	19.9	20.5	19.5	9.2	9.4	8.9	4.3	4.4	4.7
Shrink to minimum observation									
$\boldsymbol{\delta}^{GHT1}$	35.9	37.1	37.1	19.3	18.4	17.3	18.2	16.2	14.2
$\boldsymbol{\delta}^{T}$	41.5	42.0	41.4	10.5	11.4	12.6	6.0	6.2	6.8
$\boldsymbol{\delta}^{GHT2}$	41.7	4.1	−71.9	33.6	31.9	29.2	22.9	21.0	18.9
Shrink to zero									
$\boldsymbol{\delta}^{P}$	42.6	44.2	44.2	9.2	10.6	13.0	2.7	2.9	3.9
$\boldsymbol{\delta}^{CZ}$	77.5	78.5	78.6	26.2	26.8	28.5	13.8	13.9	15.1
$\boldsymbol{\delta}^{TP}$	55.7	40.4	26.4	39.5	36.5	31.7	28.4	26.8	24.6

parameters, which was not the case under the summed loss function SEL_2. This indicates that, as with the BLP under $MAXSEL_0$, these estimators can outperform the importance sampling method in a wide region of the parameter space only when the behaviour of the worst estimate of θ_i, $i = 1, \ldots, m$, is of interest and under a specific maximum component loss function.

Finally, the average risk performance of the considered methods was assessed when a larger number of Poisson means, $m = 50$, are simultaneously estimated. The same parameter settings and loss functions were used. The RARI results showed that when m increases, the average risk behaviour of the estimators follows similar patterns as in the case when $m = 10$. However, as m increases, it seems that the average risk performance of the importance sampling estimator and the BLP under the summed loss functions improves, especially in the cases that the Poisson means θ_i, $i = 1, \ldots, m$, are not concentrated close to their mean. This implies that, as $m \to \infty$, the EB methods are expected to perform as well as a hierarchical Bayes procedure which would account for the uncertainty in the hyperprior parameters. This conclusion is also supported by the fact that, unlike in the $m = 10$ case, when $m = 50$ and $\text{var}(\theta_i)$ is large, the average risk properties of the importance sampling estimator and the BLP under a summed loss function, are very close or even better than those of Morris' and Albert's methods, which are developed to better approximate the hierarchical Bayes estimator. This is demonstrated in Figure 3.2 which displays the average risk savings of the importance sampling estimator, the BLP and the Albert-type rule, when compared to the usual ML estimator, for 3 of the examined loss functions. For large m, Morris' method gives very similar results to the BLP, and therefore it is not presented in the graphs.

3.10 Summary and conclusions

In this chapter we have introduced a Bayesian structure with a single level of prior information, that being a model where the Poisson means θ_i, $i = 1, \ldots, m$, are assumed to follow independently a $\text{LN}(\mu, \sigma^2)$ prior distribution. Under this formulation, a nonclosed form arises for the posterior distribution of the parameters of interest, and therefore we seek suitable approximations for their estimation.

We consider the posterior means $\text{E}(\theta_i|\mathbf{y})$, $i = 1, \ldots, m$, of the Poisson parameters and we obtain two approximating methods. The first is a linear method which, in its general form, gives an exact linear expression of the posterior mean in the case when a conjugate prior distribution is assumed. We derive the coefficients of the linear estimator in terms of the prior mean $\xi = \text{E}(\theta_i)$ and the prior

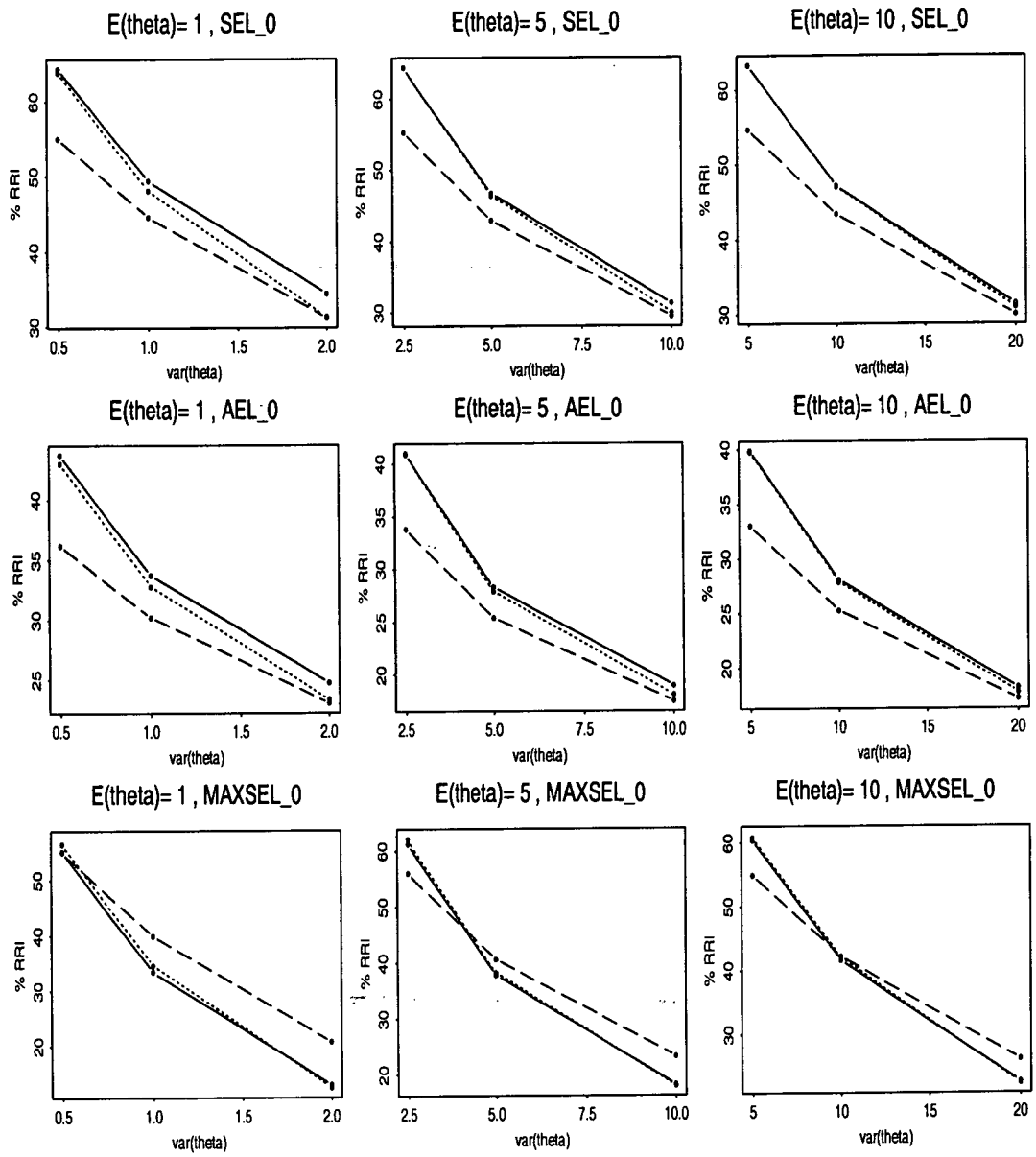

Figure 3.2: *Percentage of relative average risk improvement (RARI), in comparison to the MLE, for 3 EB methods when $m = 50$ Poisson means are estimated. The solid line (—) represents the importance sampling estimator; the dotted line (· · ·) corresponds to the BLP; and the dashed line (- - -) shows the Albert-type estimator $\boldsymbol{\delta}^A$. In each graph the RARI of the estimators is plotted versus the 3 considered values of the variance $var(\theta_i)$, for a given value of the mean $E(\theta_i)$. Each row of graphs corresponds to one of the considered loss functions, namely SEL_0, AEL_0 and $MAXSEL_0$*

variance $\phi = \text{var}(\theta_i)$ of the parameters θ_i, $i = 1, \ldots, m$, so that the resulting estimator has minimum average risk among the linear rules of the same form. The method, to which we refer as the best linear predictor (BLP), shrinks the usual estimates y_i, $i = 1, \ldots, m$, towards the prior mean $\xi = \text{E}(\theta_i)$ with a weight which depends on the prior variance $\phi = \text{var}(\theta_i)$, allowing more shrinkage under stronger prior beliefs. The second method is based on averaging simulated values drawn from the posterior distribution of θ_i, $i = 1, \ldots, m$, to estimate the posterior mean. It employs the importance sampling technique for sampling the nonclosed form of the posterior distribution, exploiting a convenient rearrangement of the latter in the logarithm scale. The resulting importance sampling estimator offers a fast and accurate nonlinear approximation to the posterior means $\text{E}(\theta_i|\mathbf{y})$, $i = 1, \ldots, m$. The implementation of both methods relies on the evaluation of the hyperparameters μ and σ^2, that is the parameters of the log-normal prior distribution of θ_i, $i = 1, \ldots, m$. As at this stage no further levels of prior knowledge are assumed, we adopt an empirical Bayes approach by estimating these parameters using the available data.

We illustrate the implementation of the developed methods for the simultaneous estimation of several Poisson means using two real data examples and two simulated data sets. The results verify that the importance sampling method yields very accurate estimates, when compared to the posterior mean derived with numerical integration. As far as the question of the shrinking direction is concerned, the importance sampling estimator seems to contradict the methods suggested until now, which adjust the MLE towards points such as zero, the minimum observation, the sample mean, the geometric mean etc. The importance sampling estimates demonstrate that the EB posterior mean shrinks the observed values towards a central value which lies between the minimum observation and the sample mean of the data.

Finally, the risk properties of the developed methods were assessed and compared to those of the estimators reviewed in Chapter 2. We considered the average risk, defined as the expectation of the frequentist risk with respect to the prior distribution of the parameters of interest. The relative average risk improvement of the estimators, when compared against the MLE, was recorded in a wide range of the parameter space. The average risk was examined under 9 different loss functions, falling in three general categories, namely the summed squared error loss, the summed absolute error loss and the maximum component squared error loss functions. The results show that the importance sampling estimator has excellent frequency properties, dominating all the considered methods in most of the examined cases. It performs well under most of the loss functions that we

employed and in the majority of the $\boldsymbol{\theta}$ regions, improving remarkably the average risk of the MLE. The only exceptions occur when a heavily weighted loss function is considered and the parameters θ_i, $i = 1, \ldots, m$, are close to zero. The BLP also gives small average risk and it compares well to the other EB rules. The latter can perform better than the importance sampling estimator and the BLP when the Poisson parameters exhibit large variation. The remaining of the examined methods produce considerable relative average risk improvement only in the region of the point towards which they shrink, and when the loss function under which they are developed is considered.

In the Bayesian formulation considered in the present chapter we only assumed one level of prior information and then proceeded with an EB approach for the estimation of the first stage hyperparameters μ and σ^2. In the remaining of this thesis, we will assume a Bayesian structure comprising more prior knowledge levels, thus forming a hierarchical Poisson model.

Chapter 4

Analytical approximations for the full hierarchical model

4.1 Introduction

In Chapter 3 we allowed only one stage of prior information, that being the prior distribution of the Poisson means $\theta_1, \theta_2, \ldots, \theta_m$. This was a log-normal prior distribution with parameters μ and σ^2 which entered the analysis either as known constants or as quantities estimated from the data. Clearly, using these parameters in the former way implies very strong prior information which normally will not be available to the statistician. Even in the case that some preliminary analyses or experiments have been conducted, the statistician might be reluctant to use this knowledge as an expression of a unique possible value for the prior parameters. On the other hand, employing an empirical Bayes approach by using the data to estimate μ and σ^2, comes under the criticism of involving the data twice in the inferential procedure. First to estimate the parameters in the prior and then to evaluate the posterior distribution via the likelihood. This procedure ignores any underlying uncertainty in the prior setting and estimation and therefore results in posterior estimates that tend to be less variable than otherwise would. Much of the opposition to the empirical Bayes methodology relies on the reasoning that it does not employ a probability distribution for the unknown parameters μ and σ^2, hence leading to an approach which some authors claim that in essence is non-Bayesian (e.g. see Deely and Lindley, 1981).

An alternative approach is to adopt a full Bayesian framework. According to this approach, at the first stage of the prior specification the population Poisson means $\theta_1, \theta_2, \ldots, \theta_m$, follow a prior distribution which depends on some unknown parameters, the so-called hyperparameters. We then assume a hyperprior distribution for the hyperparameters, forming a second stage in the prior setting, which in turn may incorporate a third prior stage and so on. This procedure creates a

hierarchical model with several stages. In practice, assigning a hyperprior distribution to the parameters at the lower stages of the model may be quite difficult, as little prior knowledge will be available at this level of the hierarchy. One way to overcome this problem is to use vague hyperprior distributions whenever there is inadequate prior information. By 'vague prior distribution', we mean a prior distribution that does not favour any specific values of the parameter of interest over the others, thus allowing the data to dominate the information contained in the prior (e.g. see Box and Tiao, 1973). This would also allow the analysis to be more objective in the case that we do not wish to be in favour of any particular set of values for the hyperparameters.

4.1.1 The hierarchical Poisson/log-normal model

We now introduce the hierarchical Poisson/log-normal model. Suppose that conditional on $\theta_1, \theta_2, \ldots, \theta_m$, the observations $y_1, y_2, \ldots, y_m$, come from independent Poisson distributions with means $\theta_1, \theta_2, \ldots, \theta_m$, respectively. We assume that the parameters θ_i , $i = 1, \ldots, m$, are independently and identically distributed according to a log-normal distribution with parameters μ and σ^2. Equivalently if we let γ_i denote the natural logarithm of θ_i, then γ_i, $i = 1, \ldots, m$, independently follow a normal $N(\mu, \sigma^2)$ distribution. At the second stage of the prior assessment we can either assume that the hyperprior parameters μ and σ^2 are both distributed according to independent vague uniform priors, or we may let σ^2 follow a scaled inverse chi-square distribution independently from μ which again is $U(-\infty, \infty)$. The first setting reflects an almost entire lack of prior information on both μ and σ^2, while the second allows some degree of prior knowledge for the population variation σ^2 to enter our analysis. Under the second specification we can still let this prior information be rather vague by selecting suitable parameters for the hyperprior scaled inverse chi-square distribution.

We have therefore described a hierarchical Poisson model with two stages of prior specification. At the lower level, we will choose the values of the hyperparameters, wherever this is needed, in such a way that they express a relative ignorance of the behaviour of the parameter at the previous stage. The model can be written as follows:

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \theta_i|\mu, \sigma^2 &\overset{iid}{\sim} \text{LN}(\mu, \sigma^2) \\ \pi(\mu) &\propto 1 \\ \pi(\sigma^2) &\propto 1 \end{aligned} \tag{4.1a}$$

where $i = 1, \ldots, m$. If we assume a scaled inverse chi-square hyperprior distribu-

tion for σ^2, with parameters ν and λ, denoted as Inv-$\chi^2(\nu, \lambda)$, the model can be expressed as

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \theta_i|\mu, \sigma^2 &\overset{iid}{\sim} \text{LN}(\mu, \sigma^2) \\ \pi(\mu) &\propto 1 \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu, \lambda) \end{aligned} \tag{4.1b}$$

where $i = 1, \ldots, m$. The probability density function of the Inv-$\chi^2(\nu, \lambda)$ is given in Appendix A. As noted above, we can equivalently express these models in terms of the logarithms of θ_i, replacing the log-normal distribution for θ_i with a normal $N(\mu, \sigma^2)$ for γ_i, $i = 1, \ldots, m$. As in Chapter 3, we let ξ and ϕ, given in (3.2), denote respectively the common mean and variance of each θ_i, $i = 1, \ldots, m$. The hierarchical Poisson/log-normal formulation that we have assumed implies a nonconjugate model. Hence, the full Bayesian analysis involves nonclosed forms of posterior densities and for any inference we will have to overcome the problem of computing intractable integrals. Two approaches to solving this problem are:

(i) Use of analytical approximations which offer a general view and understanding of the inferential procedure, although they cannot yield the final results without the partial or more extensive use of numerical or simulation methods.

(ii) Use of Monte Carlo integration methods, suitably adapted to this particular model which are implemented to obtain extended and detailed inferences.

In the present chapter the former approach will be investigated, while the latter is the subject of Chapter 5.

4.2 Analytical approximations for the conjugate model

We will attempt to estimate the unknown parameters of interest $\theta_1, \theta_2, \ldots, \theta_m$, by approximating the means of the corresponding posterior distributions. Our analytical methods are based on a normal approximation of the marginal distribution of the data. This approach might be expected to work well when the observations are large, implying large Poisson means.

We will first examine a hierarchical Poisson/Gamma model. At the first stage of the hierarchy we assume a conjugate gamma distribution for the Poisson means

with parameters $\beta\zeta$ and β. Thus, the prior probability density function of θ_i is given by

$$\pi(\theta_i|\zeta,\beta) = \frac{\beta^{\beta\zeta}\ \theta_i^{\beta\zeta-1}\exp(-\beta\theta_i)}{\Gamma(\beta\zeta)},$$

for $i = 1,\ldots,m$, where θ_i, β and ζ are positive numbers. This formulation is mathematically convenient as it will allow us to avoid using numerical integration at the early stages of our analysis. The mean of the gamma prior is ζ and its variance is proportional to the mean and equal to $\frac{\zeta}{\beta}$. The scale parameter β expresses our belief in this prior assessment. When β is large, θ_i concentrates more around its mean ζ, while as β approaches zero the prior distribution of θ_i is very dispersed around ζ, reflecting a low confidence in the prior mean of θ_i.

At the second level of the prior setting we assume that ζ and β are independent and we allow ζ to have a flat uniform hyperprior over the positive real axis. We use the reparametrisation $w = \frac{\beta}{\beta+1}$, $0 < w < 1$, and for this new hyperparameter we assume a noninformative uniform hyperprior over the unit interval. The parameter w is often referred to as the shrinkage proportion as it measures the proportion by which the conditional posterior mean of θ_i shrinks the observation y_i towards the prior mean ζ, as shown later. The model can be written as following:

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \theta_i|\zeta,\beta &\overset{iid}{\sim} \text{Ga}(\beta\zeta,\beta) \\ \pi(\zeta) \propto 1,\ \ w &= (1+\beta^{-1})^{-1} \sim U(0,1). \end{aligned} \tag{4.2}$$

Note that assuming a uniform $U(0,1)$ hyperprior for w is equivalent to assuming that the prior probability density function for β is

$$\pi(\beta) = \frac{1}{(1+\beta)^2}, \qquad 0 < \beta < \infty. \tag{4.3}$$

Leonard and Novick (1986) and Albert (1988) analyse model (4.2) in the context of two-way contingency tables and log-linear models respectively. Christiansen and Morris (1997) use a similar formulation with different parametrisation for Poisson regression modelling. They all use analytical approximations to obtain inferences about the Poisson parameters and the hyperparameters of the model. We will try to derive alternative approximations following the lines of the analysis given by Leonard (1977) for the multinomial-Dirichlet case.

Let $\boldsymbol{\theta} = (\theta_1,\theta_2,\ldots,\theta_m)^{\mathrm{T}}$ denote the vector of the Poisson means and $\mathbf{y} = (y_1,y_2,\ldots,y_m)^{\mathrm{T}}$ denote the data vector. We also denote the sampling density of

the data vector $\mathbf{y}$ given $\boldsymbol{\theta}$ by $f(\mathbf{y}|\boldsymbol{\theta})$. We finally use the notation $p(\cdot)$ and $\pi(\cdot)$ for the posterior and prior distribution respectively of any parameter of interest.

Under the conjugate gamma prior, the posterior density of the vector $\boldsymbol{\theta}$, conditional on the hyperparameters ζ and β is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y},\zeta,\beta) &\propto f(\mathbf{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}|\zeta,\beta)\ \pi(\zeta,\beta) \\ &\propto \prod_{i=1}^{m} \left\{ \left(e^{-\theta_i}\theta_i^{y_i}\right)\ \left(\theta_i^{\beta\zeta-1}e^{-\beta\theta_i}\right) \right\} \\ &= \prod_{i=1}^{m} \left\{ \theta_i^{(\beta\zeta+y_i)-1}\ e^{-(\beta+1)\theta_i} \right\}, \end{aligned} \tag{4.4}$$

i.e. it is the product of m $\mathrm{Ga}(\beta\zeta+y_i, \beta+1)$ densities. Then the conditional independence of $\theta_1, \theta_2, \ldots, \theta_m$, implies that the posterior density of each component of $\boldsymbol{\theta}$, conditional on ζ and β, is a gamma density, that is

$$\theta_i|\mathbf{y},\zeta,\beta \ \sim\ \mathrm{Ga}(\beta\zeta+y_i, \beta+1) \tag{4.5}$$

for $i = 1, \ldots, m$. Hence, the posterior expectation of θ_i conditional on ζ and β is given as

$$\mathrm{E}(\theta_i|\mathbf{y},\zeta,\beta) = \frac{\beta\zeta+y_i}{\beta+1},$$

and this can be written in the linear form

$$\begin{aligned} \mathrm{E}(\theta_i|\mathbf{y},\zeta,\beta) &= (1-w)y_i\ +\ w\zeta \\ &= y_i\ +\ w(\zeta-y_i), \end{aligned} \tag{4.6}$$

where $w = (1+\beta^{-1})^{-1}$, $0 < w < 1$. The linearity of the above posterior mean is a direct consequence of the conjugacy in the first stage of the hierarchical model. From this linear form it can be seen that the posterior mean is a shrinkage estimator, with w giving the proportion of the distance between the MLE, that is the observation y_i, and the prior mean ζ, in which y_i is 'pulled-in' towards the direction of ζ. Clearly, if we solve (4.6) for w we have that

$$w = \frac{\mathrm{E}(\theta_i|\mathbf{y},\zeta,\beta) - y_i}{\zeta - y_i}, \qquad 0 < w < 1. \tag{4.7}$$

To obtain the unconditional posterior expectations of the Poisson means θ_i we must average the conditional expectation (4.6) with respect to the posterior distribution of the hyperparameters ζ and β. The latter is given by

$$p(\zeta,\beta|\mathbf{y}) \propto f(\mathbf{y}|\zeta,\beta)\ \pi(\zeta,\beta), \tag{4.8}$$

where $f(\mathbf{y}|\zeta,\beta)$ denotes the marginal distribution of the data and $\pi(\zeta,\beta)$ is the joint prior of the hyperparameters.

Exact derivation of $\mathrm{E}(\theta_i|\mathbf{y})$

We first consider the exact marginal density of the data. This can be calculated using the joint distribution of $\mathbf{y}$ and $\boldsymbol{\theta}$ given ζ and β, that is

$$\begin{aligned} f(\mathbf{y},\boldsymbol{\theta}|\zeta,\beta) &= f(\mathbf{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}|\zeta,\beta) \\ &\propto \prod_{i=1}^{m}\left\{\left(\frac{1}{y_i!}\ e^{-\theta_i}\theta_i^{y_i}\right)\ \left(\theta_i^{\beta\zeta-1}e^{-\beta\theta_i}\right)\right\} \\ &= \prod_{i=1}^{m}\left\{\frac{1}{y_i!}\ \theta_i^{(\beta\zeta+y_i)-1}\ e^{-(\beta+1)\theta_i}\right\}, \end{aligned}$$

defined for $\mathbf{y}$ and $\boldsymbol{\theta}$ being m-dimensional vectors of nonnegative integers and positive reals respectively. Then, the marginal density for $\mathbf{y}$ can be obtained by integrating the parameter vector $\boldsymbol{\theta}$ out of the above joint density, i.e.

$$\begin{aligned} f(\mathbf{y}|\zeta,\beta) &= \int_{\Theta} f(\mathbf{y},\boldsymbol{\theta}|\zeta,\beta)d\boldsymbol{\theta} \\ &\propto \int_{\Theta}\prod_{i=1}^{m}\left\{\frac{1}{y_i!}\ \theta_i^{(\beta\zeta+y_i)-1}\ e^{-(\beta+1)\theta_i}\right\}d\boldsymbol{\theta} \\ &\propto \prod_{i=1}^{m}\left\{\frac{1}{y_i!}\ \int_0^{\infty}\theta_i^{(\beta\zeta+y_i)-1}\ e^{-(\beta+1)\theta_i}d\theta_i\right\}, \end{aligned}$$

where Θ denotes the domain of the $\boldsymbol{\theta}$ vector, that is the m-dimensional space of positive reals. The integrand in the above expression is proportional to a $\mathrm{Ga}(\beta\zeta+y_i,\beta+1)$ density and therefore, ignoring the constant of proportionality we can write the probability function of the observations as

$$\begin{aligned} f(\mathbf{y}|\zeta,\beta) &\propto \prod_{i=1}^{m}\left\{\frac{1}{y_i!}\ \frac{\Gamma(\beta\zeta+y_i)}{(\beta+1)^{\beta\zeta+y_i}}\right\} \\ &\propto \prod_{i=1}^{m}\left\{\frac{\Gamma(y_i+\beta\zeta)}{y_i!}\ \left(\frac{1}{\beta+1}\right)^{y_i}\right\}, \end{aligned}$$

for $\mathbf{y} = (y_1,\ldots,y_m)^{\mathrm{T}}$, where y_i, $i = 1,\ldots,m$, are nonnegative integers. The above expression shows that the resulting distribution of the data vector is a product of negative binomial densities with parameters $\beta\zeta$ and β, or equivalently, the marginal distribution of y_i, obtained by summing the rest of the observations, is negative binomial with the same parameters, i.e.

$$Y_i \overset{iid}{\sim} \text{Neg.Bin.}(\beta\zeta,\beta), \qquad i = 1,\ldots,m, \tag{4.9}$$

with mean and variance given as

$$\mathrm{E}(Y_i) = \zeta, \qquad \mathrm{var}(Y_i) = (1+\beta^{-1})\zeta. \tag{4.10}$$

We can now return to the joint posterior distribution of ζ and β in (4.8). Using the marginal distribution for $\mathbf{y}$ derived above and the prior for (ζ, β) specified in model (4.2) and in (4.3) we obtain that

$$p(\zeta, \beta|\mathbf{y}) \propto (1+\beta)^{-2} \prod_{i=1}^{m} \left\{ \binom{y_i + \beta\zeta - 1}{\beta\zeta - 1} \left(\frac{\beta}{\beta+1}\right)^{\beta\zeta} \left(\frac{1}{\beta+1}\right)^{y_i} \right\}. \quad (4.11)$$

Therefore the unconditional posterior mean of θ_i will be given by

$$\mathrm{E}(\theta_i|\mathbf{y}) = \int_0^\infty \int_0^\infty \mathrm{E}(\theta_i|\mathbf{y}, \zeta, \beta)\, p(\zeta, \beta|\mathbf{y})\, d\zeta d\beta \quad (4.12)$$

with the conditional mean $\mathrm{E}(\theta_i|\mathbf{y}, \zeta, \beta)$ given in (4.6) and $p(\zeta, \beta|\mathbf{y})$ given in (4.11). In practice, due to the form of the joint posterior of ζ and β, the evaluation of the above double integral is a complicated task which would not allow us to obtain any further analytical results without the use of numerical methods. This suggests that employing a simpler form for $p(\zeta, \beta|\mathbf{y})$ would facilitate the calculation of the posterior expectation of θ_i, $i = 1, \ldots, m$.

4.2.1 Approximate $\mathrm{E}(\theta_i|\mathbf{y})$ using a normal approximation

We will consider a normal approximation to the marginal distribution of the data, allowing the random variables $Y_1, Y_2, \ldots, Y_m$, to follow independent normal distributions having the exact first two moments as these are given by the negative binomial distribution in (4.10). We will also use the reparametrisation $w = (1+\beta^{-1})^{-1}$ introduced earlier in this section. Then approximately

$$Y_i \overset{iid}{\sim} N(\zeta, w^{-1}\zeta), \qquad i = 1, \ldots, m. \quad (4.13)$$

Using this approximation we can derive the approximate posterior distribution of the hyperparameters ζ and w from

$$p(\zeta, w|\mathbf{y}) \propto f(\mathbf{y}|\zeta, w)\, \pi(\zeta, w),$$

where $f(\mathbf{y}|\zeta, w)$ is the approximate marginal multivariate normal density following directly from (4.13) and $\pi(\zeta, w)$ is the prior for the hyperparameters defined in model (4.2). Hence,

$$\begin{aligned} p(\zeta, w|\mathbf{y}) &\propto \prod_{i=1}^{m} \left\{ (w^{-1}\zeta)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\, w\zeta^{-1}(y_i - \zeta)^2 \right] \right\} \\ &\propto \zeta^{-\frac{m}{2}} w^{\frac{m}{2}} \exp\left\{ -\frac{1}{2}\, w\, \frac{\sum_{i=1}^{m}(y_i - \zeta)^2}{\zeta} \right\}, \end{aligned} \quad (4.14)$$

where $0 < w < 1$ and $0 < \zeta < \infty$. Equation (4.14) provides a simpler form for the posterior distribution of the hyperparameters. To obtain the unconditional expectation of θ_i we first notice that from (4.12) and under the w reparametrisation, we can write

$$\mathrm{E}(\theta_i|\mathbf{y}) = \int_0^\infty \int_0^1 \mathrm{E}(\theta_i|\mathbf{y}, \zeta, w)\ p(\zeta, w|\mathbf{y})\ dw d\zeta$$

$$= \int_0^\infty \left\{ \int_0^1 \mathrm{E}(\theta_i|\mathbf{y}, \zeta, w)\ p(w|\zeta, \mathbf{y}) dw \right\} p(\zeta|\mathbf{y}) d\zeta$$

and as the integral in the brackets equals the posterior expectation of θ_i conditional only on ζ, we obtain

$$\mathrm{E}(\theta_i|\mathbf{y}) = \int_0^\infty \mathrm{E}(\theta_i|\zeta, \mathbf{y})\ p(\zeta|\mathbf{y}) d\zeta. \tag{4.15}$$

The posterior expectation of θ_i conditional on ζ, as it appears in the integrand of the above expression, is given as

$$\mathrm{E}(\theta_i|\zeta, \mathbf{y}) = \int_0^1 \mathrm{E}(\theta_i|\zeta, w, \mathbf{y})\ p(w|\zeta, \mathbf{y})\ dw$$

and using (4.6)

$$\mathrm{E}(\theta_i|\zeta, \mathbf{y}) = \int_0^1 \{(1-w)y_i + w\zeta\} p(w|\zeta, \mathbf{y})\ dw$$

$$= \{1 - \mathrm{E}(w|\zeta, \mathbf{y})\}\ y_i\ +\ \mathrm{E}(w|\zeta, \mathbf{y})\ \zeta. \tag{4.16}$$

Now, expression (4.14) implies that, conditional on ζ, the posterior of w is a $\mathrm{Ga}\left(\frac{m}{2}+1, \frac{1}{2}\frac{\sum_{i=1}^m (y_i-\zeta)^2}{\zeta}\right)$ density truncated to the $(0,1)$ interval, i.e.

$$p(w|\zeta, \mathbf{y}) \propto w^{\frac{m}{2}} \exp\left\{-\frac{1}{2}\ w\ \frac{\sum_{i=1}^m (y_i-\zeta)^2}{\zeta}\right\}, \qquad 0 < w < 1. \tag{4.17}$$

Therefore, the posterior expectation of w conditional on ζ will be equal to

$$\mathrm{E}(w|\zeta, \mathbf{y}) = c^{-1} \int_0^1 w\ w^{\frac{m}{2}} \exp\left\{-\frac{1}{2}\ w\ \frac{\sum_{i=1}^m (y_i-\zeta)^2}{\zeta}\right\} dw,$$

where c is the normalising constant for the posterior density of w in (4.17). If we let

$$X^2 = \frac{\sum_{i=1}^m (y_i-\zeta)^2}{\zeta}, \tag{4.18}$$

the conditional mean of w above becomes

$$\mathrm{E}(w|\zeta,\mathbf{y}) = \frac{\int_0^1 w^{\frac{1}{2}(m+2)}\exp\left\{-\frac{1}{2}\ wX^2\right\}dw}{\int_0^1 w^{\frac{1}{2}m}\exp\left\{-\frac{1}{2}\ wX^2\right\}dw}$$

$$= \frac{J_{m+2}(X^2)}{J_m(X^2)} \tag{4.19}$$

where $J_{m+k}(X^2)$ is a function of X^2 defined as

$$J_{m+k}(X^2) = \int_0^1 t^{\frac{1}{2}(m+k)}\exp\left\{-\frac{1}{2}\ tX^2\right\}dt. \tag{4.20}$$

The posterior of ζ in the integrand of (4.15) can be derived if we integrate w out of their joint posterior distribution. From (4.14) we have

$$p(\zeta|\mathbf{y}) = \int_0^1 p(\zeta,w|\mathbf{y})\ dw$$

$$\propto \zeta^{-\frac{m}{2}}\ \int_0^1 w^{\frac{1}{2}m}\exp\left\{-\frac{1}{2}\ wX^2\right\}dw$$

$$\propto \zeta^{-\frac{m}{2}}\ J_m(X^2), \tag{4.21}$$

with X^2 and $J_m(X^2)$ defined in (4.18) and (4.20) respectively. We can now obtain the unconditional posterior mean of θ_i, by combining equations (4.15), (4.16), (4.19) and (4.21). This will give

$$\mathrm{E}(\theta_i|\mathbf{y}) = \int_0^\infty [\{1-\mathrm{E}(w|\zeta,\mathbf{y})\}y_i\ +\ \mathrm{E}(w|\zeta,\mathbf{y})\zeta]\ p(\zeta|\mathbf{y})\ d\zeta$$

$$\propto \int_0^\infty \zeta^{-\frac{m}{2}}J_m(X^2)\left\{1-\frac{J_{m+2}(X^2)}{J_m(X^2)}\right\}y_i d\zeta$$

$$+\ \int_0^\infty \zeta^{-\frac{m}{2}+1}J_m(X^2)\left\{\frac{J_{m+2}(X^2)}{J_m(X^2)}\right\}d\zeta$$

$$= \left[\int_0^\infty \zeta^{-\frac{m}{2}}\{J_m(X^2)-J_{m+2}(X^2)\}d\zeta\right]y_i$$

$$+\ \int_0^\infty \zeta^{-\frac{m}{2}+1}J_{m+2}(X^2)d\zeta. \tag{4.22}$$

4.2.2 Approximate $\mathbf{E}(\theta_i|\mathbf{y})$ based on the χ^2 statistic

The evaluation of the posterior mean in (4.22) involves the computation of mathematically intractable integrals, including two-dimensional ones, which need to be carried out using numerical integration techniques. As this can be tedious and time consuming, we suggest an alternative approach employing the same normal approximation for the marginal distribution of the data and also an approximation based on the chi-square statistic. If, as earlier, we assume that marginally the data are approximately distributed as

$$Y_i \overset{iid}{\sim} N(\zeta, w^{-1}\zeta), \qquad i = 1, \ldots, m,$$

it follows that the quadratic form $\frac{\sum_{i=1}^{m}(y_i-\zeta)^2}{w^{-1}\zeta}$ conditional on ζ and w has an approximate chi-square distribution with m degrees of freedom, as shown by Stuart and Ord (1994). Using the notation in (4.18) we can write

$$wX^2|\zeta, w \;\dot{\sim}\; \chi^2_m. \tag{4.23}$$

We then replace ζ in X^2 by its moment estimate, that is the sample mean of the data $\bar{y}$. We let $\widetilde{X}^2$ denote the new statistic, i.e. we write

$$\widetilde{X}^2 = \frac{\sum_{i=1}^{m}(y_i - \bar{y})^2}{\bar{y}}\,. \tag{4.24}$$

The estimation of ζ in X^2 results in a decrease of the degrees of freedom in the chi-square approximation (4.23) which now becomes

$$w\widetilde{X}^2|w \;\dot{\sim}\; \chi^2_{m-1}. \tag{4.25}$$

This means that given w the approximate probability density function of $w\widetilde{X}^2$ is

$$\pi\left(w\widetilde{X}^2|w\right) \propto \left(w\widetilde{X}^2\right)^{\frac{m-1}{2}-1} \exp\left\{-\frac{1}{2}\, w\widetilde{X}^2\right\}$$

and hence the conditional distribution of $\widetilde{X}^2$ is approximated by

$$\pi\left(\widetilde{X}^2|w\right) \propto \left(\widetilde{X}^2\right)^{\frac{m-1}{2}-1} \exp\left\{-\frac{1}{2}\, w\widetilde{X}^2\right\},$$

where $\widetilde{X}^2 > 0$, and $0 < w < 1$. This is a $\mathrm{Ga}(\frac{m-1}{2}, \frac{w}{2})$ density and allowing for the normalising constant we can write

$$\pi\left(\widetilde{X}^2|w\right) = \frac{\left(\frac{w}{2}\right)^{\frac{m-1}{2}}}{\Gamma\left(\frac{m-1}{2}\right)} \left(\widetilde{X}^2\right)^{\frac{m-1}{2}-1} \exp\left\{-\frac{1}{2}\, w\widetilde{X}^2\right\}.$$

Note that $\widetilde{X}^2$ contains all the available information about the data. Thus, if we ignore the information about w possessed in the data and not included in the density above, the likelihood function of w will be given by

$$L(w|\mathbf{y}) \propto w^{\frac{m-1}{2}} \exp\left\{-\frac{1}{2}\, w\widetilde{X}^2\right\}. \tag{4.26}$$

The prior distribution of the hyperparameter w is specified in model (4.2) as uniform over the unit interval, and therefore we can now calculate the approximate posterior density of the hyperparameter w, using Bayes' formula to obtain

$$\begin{aligned} p(w|\mathbf{y}) &\propto L(w|\mathbf{y})\ \pi(w) \\ &\propto w^{\frac{m-1}{2}} \exp\left\{-\frac{1}{2}\, w\widetilde{X}^2\right\}, \end{aligned} \tag{4.27}$$

with $0 < w < 1$.

Comparing equations (4.17) and (4.27) we notice that estimating the hyperprior mean ζ by $\bar{y}$ in the latter, rather than conditioning on it in the former, results in a similar approximate posterior density for w. The unconditional density in (4.27) is Ga $\left(\frac{m-1}{2}+1, \frac{1}{2}\widetilde{X}^2\right)$ truncated to $(0,1)$, which is an adjusted form of the previously conditional on ζ Gamma $\left(\frac{m}{2}+1, \frac{1}{2}X^2\right)$ density, reflecting the corresponding decrease in the degrees of freedom in the chi-square distribution of $w\widetilde{X}^2$ in (4.25).

We return now to the conditional posterior expectation of θ_i, which was given earlier as

$$\mathrm{E}(\theta_i|\mathbf{y}, \zeta, \beta) = (1-w)y_i \ + \ w\zeta.$$

If we average with respect to the posterior distribution of w and ζ, the unconditional mean is given by

$$\begin{aligned} \mathrm{E}(\theta_i|\mathbf{y}) &= \mathrm{E}\{(1-w)|\mathbf{y}\}\ y_i \ + \ \mathrm{E}(w\zeta|\mathbf{y}) \\ &= (1-w^*)\ y_i \ + \ w^* \left\{\frac{\mathrm{E}(w\zeta|\mathbf{y})}{w^*}\right\}, \end{aligned} \tag{4.28}$$

where w^* denotes the posterior mean of w, i.e

$$w^* = \mathrm{E}(w|\mathbf{y}) \tag{4.29}$$

and $\frac{\mathrm{E}(w\zeta|\mathbf{y})}{w^*}$ is the quantity towards which the posterior mean shrinks the observed value y_i. Alternatively, if we estimate ζ in $\mathrm{E}(\theta_i|\mathbf{y}, \zeta, \beta)$ by the data sample mean, we need only average over w to obtain

$$\mathrm{E}(\theta_i|\mathbf{y}) = (1-w^*)\ y_i \ + \ w^*\ \bar{y}. \tag{4.30}$$

We now use the posterior density of w derived earlier to obtain an approximate expression for w^*. Notice that

$$\begin{aligned} w^* = \mathrm{E}(w|\mathbf{y}) &= \int_0^1 w\ p(w|\mathbf{y})\ dw \\ &\doteq c^{-1} \int_0^1 w\ w^{\frac{m-1}{2}} \exp\left\{-\frac{1}{2}\ w\widetilde{X}^2\right\}\ dw, \end{aligned}$$

where $c = \int_0^1 w^{\frac{m-1}{2}} \exp\left\{-\frac{1}{2}\ w\widetilde{X}^2\right\} dw$ is the normalising constant. Hence, approximately

$$\begin{aligned} w^* &= \frac{\int_0^1 w^{\frac{m+1}{2}} \exp\left\{-\frac{1}{2}\ w\widetilde{X}^2\right\}\ dw}{\int_0^1 w^{\frac{m-1}{2}} \exp\left\{-\frac{1}{2}\ w\widetilde{X}^2\right\}\ dw} \\ &= \frac{J_{m+1}(\widetilde{X}^2)}{J_{m-1}(\widetilde{X}^2)}, \end{aligned} \tag{4.31}$$

with $J_{m+1}(\widetilde{X}^2)$ introduced earlier in (4.20). Substituting w^* in (4.30) will yield the approximate posterior expectations of the Poisson parameters θ_i. In doing so we need only perform two numerical integrations.

We can further simplify the form of w^* by using the incomplete gamma function and a simple integration by parts result.

Lemma 4.1. *Let $\gamma(a, x)$ denote the incomplete gamma function defined as*

$$\gamma(a, x) = \int_0^x u^{a-1} e^{-u} du. \tag{4.32}$$

Then,

$$\gamma(a+1, x) = a\ \gamma(a, x)\ -\ x^a e^{-x}. \tag{4.33}$$

Proof. The integration by parts rule implies that

$$\begin{aligned} \gamma(a, x) &= \frac{u^a}{a}\ e^{-u}\Big|_0^x\ +\ \int_0^x \frac{u^a}{a}\ e^{-u}\ du \\ &= \frac{1}{a}\ x^a e^{-x}\ +\ \frac{1}{a}\ \gamma(a+1, x) \end{aligned}$$

which can be rearranged to give (4.33). □

Returning to w^* we first notice that we can express the function $J_{m+k}(\widetilde{X}^2)$ in terms of the incomplete gamma function in (4.32). Applying the transformation

$u = \frac{1}{2}\, t\widetilde{X}^2$ we obtain

$$J_{m+k}(\widetilde{X}^2) = \int_0^1 t^{\frac{1}{2}(m+k)} \exp\left\{-\frac{1}{2}\, t\widetilde{X}^2\right\} dt$$

$$= 2^{\frac{1}{2}(m+k)+1} \left(\widetilde{X}^2\right)^{-\frac{1}{2}(m+k)-1} \int_0^{\frac{1}{2}\widetilde{X}^2} u^{\frac{1}{2}(m+k)} e^{-u} du$$

$$= 2^{\frac{1}{2}(m+k)+1} \left(\widetilde{X}^2\right)^{-\frac{1}{2}(m+k)-1} \gamma\left\{\frac{1}{2}(m+k)+1, \frac{1}{2}\widetilde{X}^2\right\}.$$

Therefore, from (4.31), w^* can be written as

$$w^* = \frac{2^{\frac{1}{2}(m+3)} \left(\widetilde{X}^2\right)^{-\frac{1}{2}(m+3)} \gamma\left\{\frac{1}{2}(m+1)+1, \frac{1}{2}\widetilde{X}^2\right\}}{2^{\frac{1}{2}(m+1)} \left(\widetilde{X}^2\right)^{-\frac{1}{2}(m+1)} \gamma\left\{\frac{1}{2}(m+1), \frac{1}{2}\widetilde{X}^2\right\}}$$

$$= \left(\frac{1}{2}\widetilde{X}^2\right)^{-1} \frac{\gamma\left\{\frac{1}{2}(m+1)+1, \frac{1}{2}\widetilde{X}^2\right\}}{\gamma\left\{\frac{1}{2}(m+1), \frac{1}{2}\widetilde{X}^2\right\}}, \tag{4.34}$$

and if we apply (4.33) from Lemma 4.1 we obtain

$$w^* = \left(\frac{1}{2}\widetilde{X}^2\right)^{-1} \frac{\frac{1}{2}(m+1)\, \gamma\left\{\frac{1}{2}(m+1), \frac{1}{2}\widetilde{X}^2\right\} - \left(\frac{1}{2}\widetilde{X}^2\right)^{\frac{1}{2}(m+1)} e^{-\frac{1}{2}\widetilde{X}^2}}{\gamma\left\{\frac{1}{2}(m+1), \frac{1}{2}\widetilde{X}^2\right\}}$$

$$= (m+1)\widetilde{X}^{-2} - \frac{\left(\frac{1}{2}\widetilde{X}^2\right)^{\frac{1}{2}(m-1)} e^{-\frac{1}{2}\widetilde{X}^2}}{\gamma\left\{\frac{1}{2}(m+1), \frac{1}{2}\widetilde{X}^2\right\}}. \tag{4.35}$$

The last expression of w^* implies that we can approximately evaluate the posterior mean $\mathrm{E}(\theta_i|\mathbf{y})$ from (4.30) performing only one numerical integration. We have therefore provided a simpler alternative method for the estimation of the Poisson means of the hierarchical model. Furthermore, form (4.35) allows a direct comparison with some of the EB approaches presented in Chapter 2. Remember that for the Leonard (1976) EB estimator in (2.32), the shrinkage coefficient C_L is given in terms of $\widetilde{X}^2 = \frac{\sum_{i=1}^m (y_i - \bar{y})^2}{\bar{y}}$ by

$$C_L = \min\left\{(m-1)\widetilde{X}^{-2}, 1\right\}, \tag{4.36}$$

and Morris (1983) assumes a multiplicative adjustment term $\frac{m-3}{m-1}$ to the above shrinkage proportion, to obtain the coefficient C_M in (2.35). The difference be-

tween w^* and the shrinking coefficient C, which can be either C_L or C_M, is

$$w^* - C = \left(\frac{1}{2}\widetilde{X}^2\right)^{-1} \frac{\frac{\lambda}{2}(m+1)\,\gamma\left\{\frac{1}{2}(m+1), \frac{1}{2}\widetilde{X}^2\right\} - \left(\frac{1}{2}\widetilde{X}^2\right)^{\frac{1}{2}(m+1)} e^{-\frac{1}{2}\widetilde{X}^2}}{\gamma\left\{\frac{1}{2}(m+1), \frac{1}{2}\widetilde{X}^2\right\}},$$

where λ equals 2 or 4 for Leonard's or Morris' estimator respectively. This suggests, as shown in Figure 4.1, that the EB estimates provide more radical shrinkages for relatively small values of $\widetilde{X}^2$, that is when the observations are close to each other. On the other hand, as $\widetilde{X}^2$ increases, indicating an increase in the dispersion of the data, the shrinking coefficients converge to the same value. The sign and magnitude of the difference between w^* and C as $\widetilde{X}^2$ approaches its limits, is explained by the asymptotic behaviour of each coefficient in the limiting cases.

Clearly, as a consequence of the restriction in (4.36), the EB shrinkage coefficient lies within the interval $(0,1)$, or $(0, \frac{m-3}{m-1})$ for Morris' adjusted version. The parameter w^* never exceeds 1, unlike C_L or C_M in the EB estimators to which we need to impose certain constraints in order to lie in the unit interval. In fact w^* lies in the interval $(0, \frac{m+1}{m+3})$. The lower bound is attained when $\widetilde{X}^2 \to \infty$, since from (4.34) we have

$$\begin{aligned}\lim_{\widetilde{X}^2\to\infty} w^* &= \lim_{\widetilde{X}^2\to\infty}\left\{\left(\frac{1}{2}\widetilde{X}^2\right)^{-1}\frac{\int_0^{\frac{1}{2}\widetilde{X}^2} u^{\frac{1}{2}(m+1)}e^{-u}du}{\int_0^{\frac{1}{2}\widetilde{X}^2} u^{\frac{1}{2}(m-1)}e^{-u}du}\right\}\\ &= 0 \times \frac{\Gamma\left\{\frac{1}{2}(m+1)+1\right\}}{\Gamma\left\{\frac{1}{2}(m+1)\right\}} = 0 \times \frac{1}{2}(m+1) = 0.\end{aligned}$$

In the limiting case that $\widetilde{X}^2$ approaches the origin we have

$$\begin{aligned}\lim_{\widetilde{X}^2\to 0} w^* &= \lim_{\widetilde{X}^2\to 0}\left\{\left(\frac{1}{2}\widetilde{X}^2\right)^{-1}\frac{\int_0^{\frac{1}{2}\widetilde{X}^2} u^{\frac{1}{2}(m+1)}e^{-u}du}{\int_0^{\frac{1}{2}\widetilde{X}^2} u^{\frac{1}{2}(m-1)}e^{-u}du}\right\}\\ &= \lim_{\widetilde{X}^2\to 0}\left\{\frac{\int_0^{\widetilde{X}^2}\left(\frac{v}{2}\right)^{\frac{1}{2}(m+1)}e^{-\frac{v}{2}}dv}{\frac{1}{2}\widetilde{X}^2\int_0^{\widetilde{X}^2}\left(\frac{v}{2}\right)^{\frac{1}{2}(m+1)}e^{-\frac{v}{2}}dv}\right\},\end{aligned}$$

which results in an indeterminate $\frac{0}{0}$ form. We proceed applying l'Hospital's rule twice to obtain

$$\begin{aligned}\lim_{\widetilde{X}^2\to 0} w^* &= \lim_{\widetilde{X}^2\to 0}\left\{\frac{\frac{1}{2}(m+1) - \left(\frac{1}{2}\widetilde{X}^2\right)^2}{\frac{1}{2}(m+3) - \left(\frac{1}{2}\widetilde{X}^2\right)^2}\right\}\\ &= \frac{m+1}{m+3}.\end{aligned}$$

Therefore, for small $\widetilde{X}^2$ values the shrinkage proportion approaches the ratio $\frac{m+1}{m+3}$ which asymptotically tends to 1 as m gets large.

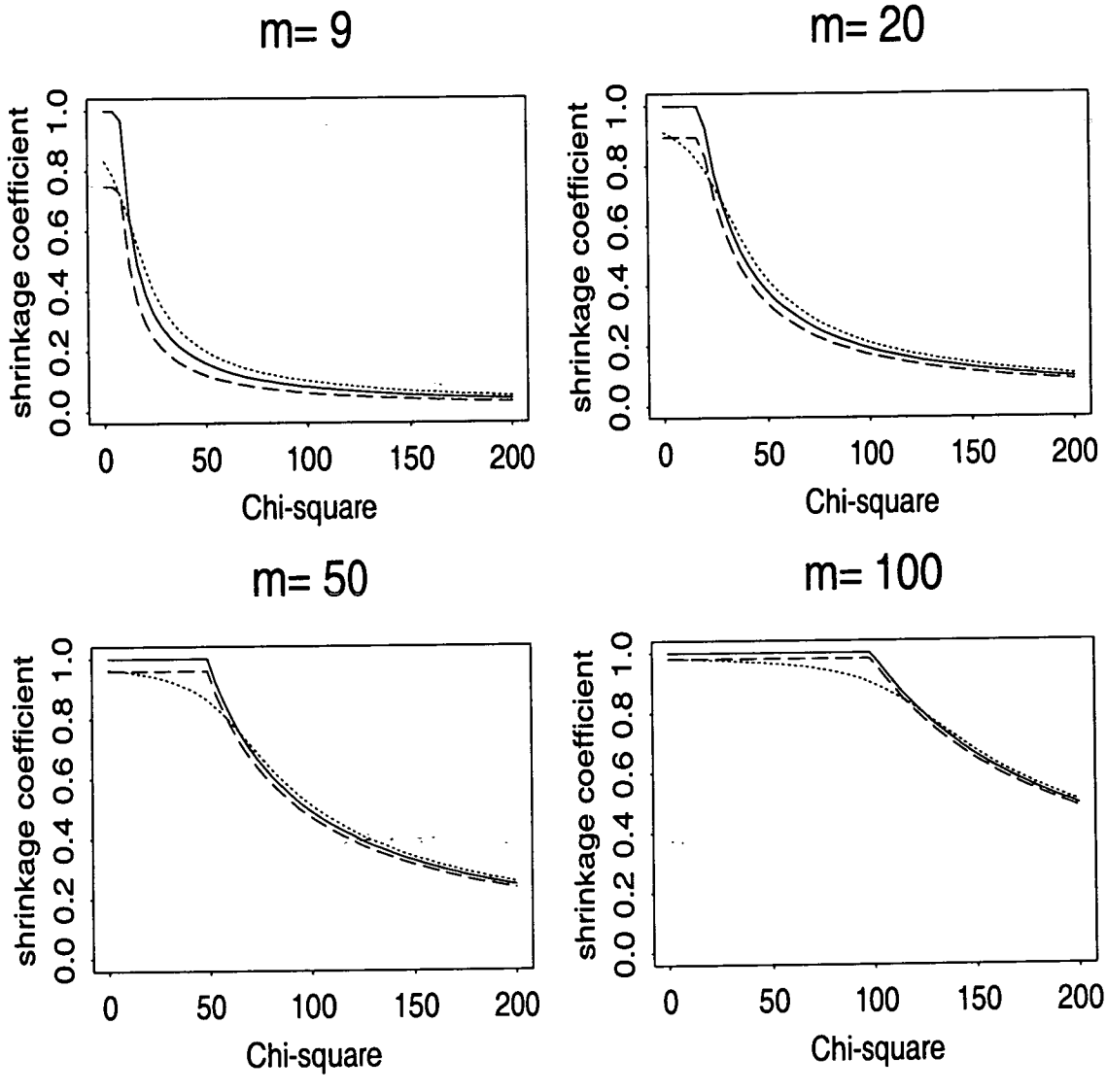

Figure 4.1: Comparison of the three shrinking coefficients for different values of the number of Poisson means (m): Leonard's C_L (—); w^* ($\cdots$); and Morris' C_M (- - -).

4.3 Analytical approximations for the Poisson/log-normal model

We will now consider an approximate Bayesian analysis for the Poisson/log-normal model (4.1a). The analysis is based again on a normal approximation to the marginal density of the data and the procedure is similar to that of the preceding section. However, the absence of conjugacy in this formulation implies that supplying analytical forms of inference is even more difficult than before, and that the use of numerical methods will be required at an earlier stage of the analysis.

According to model (4.1a), we replace the first stage gamma prior of the Poisson means θ_i, $i = 1, \ldots, m$ in the preceding section, with a log-normal distribution, having μ and σ^2 as parameters. As mentioned earlier this is equivalent to allowing the natural logarithms of the parameters $\theta_1, \theta_2, \ldots, \theta_m$, to be independently and identically normally distributed with common mean μ and common variance σ^2, that is $\gamma_i = \log(\theta_i) \sim N(\mu, \sigma^2)$. The hyperparameter σ^2 expresses the variability within each $\gamma_i = \log(\theta_i)$, and thus large values of it indicate a low belief in the prior estimate μ of γ_i. At the second stage of the prior hierarchy we assume independent flat uniform hyperprior distributions for μ and σ^2 over their domains, $(-\infty, \infty)$ and $(0, \infty)$ respectively.

Using the notation introduced earlier in this chapter, model (4.1a) implies that the posterior density for the vector of the Poisson means $\boldsymbol{\theta}$, conditional on μ and σ^2 is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, \mu, \sigma^2) &\propto f(\mathbf{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}|\mu, \sigma^2) \times \pi(\mu, \sigma^2) \\ &\propto \prod_{i=1}^{m} \left[\frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!}\ \frac{1}{\theta_i} \exp\left\{ -\frac{1}{2}\sigma^{-2}(\log\theta_i - \mu)^2 \right\} \right] \\ &\propto \prod_{i=1}^{m} \left(\theta_i^{y_i - 1}\right)\ \exp\left[\sum_{i=1}^{m} \left\{ -\theta_i - \frac{1}{2}\sigma^{-2}(\log\theta_i - \mu)^2 \right\} \right], \end{aligned} \tag{4.37}$$

where $\theta_i > 0$, $i = 1, \ldots, m$. In terms of $\gamma_i = \log(\theta_i)$, we can write

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{y}, \mu, \sigma^2) &\propto f(\mathbf{y}|\boldsymbol{\gamma})\ \pi(\boldsymbol{\gamma}|\mu, \sigma^2)\ \pi(\mu, \sigma^2) \\ &\propto \prod_{i=1}^{m} \left[\exp\left\{\gamma_i y_i - e^{\gamma_i}\right\}\ \exp\left\{ -\frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2 \right\} \right] \\ &\propto \exp\left[\sum_{i=1}^{m} \left\{ \gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2 \right\} \right], \end{aligned} \tag{4.38}$$

where $-\infty < \gamma_i < \infty$, $i = 1, \ldots, m$. Clearly the posterior density in (4.37) or (4.38) does not assume any known form. Therefore the calculation of the

unconditional posterior expectation relies on integration of a nonclosed function of $\boldsymbol{\theta}$, in addition to averaging with respect to the posterior distribution of the hyperparameters μ and σ^2, i.e.

$$\mathrm{E}(\boldsymbol{\theta}|\mathbf{y}) = \int \boldsymbol{\theta}\; p(\boldsymbol{\theta}|\mathbf{y})\; d\boldsymbol{\theta} = \iiint \boldsymbol{\theta}\; p(\boldsymbol{\theta}|\mathbf{y}, \mu, \sigma^2)\; p(\mu, \sigma^2|\mathbf{y})\; d\mu\; d\sigma^2 d\boldsymbol{\theta}.$$

4.3.1 A linear approximation to $\mathrm{E}(\theta_i|\mathbf{y})$

We will now attempt to simplify the above calculation. According to the model, the prior mean and variance of θ_i, $i = 1, \ldots, m$, are given by ξ and ϕ in (3.2). In Chapter 3 we derived a linear approximation to the conditional posterior mean of θ_i given ξ and ϕ, that being the BLP. If we let $\widetilde{\mathrm{E}}(\theta_i|\mathbf{y}, \xi, \phi)$ denote the approximated conditional posterior mean, we can write

$$\widetilde{\mathrm{E}}(\theta_i|\mathbf{y}, \xi, \phi) = (1 - \rho)y_i \;+\; \rho\xi \tag{4.39}$$

where, from (3.15)

$$\rho = \frac{\mathrm{E}(\theta_i)}{\mathrm{var}(\theta_i) + \mathrm{E}(\theta_i)} = \frac{\xi}{\phi + \xi}\,. \tag{4.40}$$

Notice that the above result holds exactly for the conjugate Poisson/Gamma model, providing the exact linear form (4.6) for the posterior mean.

To obtain the unconditional approximate posterior expectation of θ_i, denoted hereafter as $\widetilde{\mathrm{E}}(\theta_i|\mathbf{y})$, we need to average (4.39) with respect to the posterior density of the hyperparameters ξ and ϕ, i.e.

$$\begin{aligned}
\widetilde{\mathrm{E}}(\theta_i|\mathbf{y}) &= \int_0^\infty \int_0^\infty \widetilde{\mathrm{E}}(\theta_i|\mathbf{y}, \xi, \phi)\; p(\xi, \phi|\mathbf{y})\; d\xi d\phi \\
&= \int_0^\infty \int_0^\infty \{(1 - \rho)y_i + \rho\xi\}\; p(\xi, \phi|\mathbf{y})\; d\xi d\phi \\
&= \mathrm{E}\{(1 - \rho)|\mathbf{y}\}y_i \;+\; \mathrm{E}(\rho\xi|\mathbf{y}) \qquad (4.41) \\
&= (1 - \rho^*)y_i \;+\; \rho^* \frac{\mathrm{E}(\rho\xi|\mathbf{y})}{\mathrm{E}(\rho|\mathbf{y})} \qquad (4.42)
\end{aligned}$$

where ρ^* is the posterior expectation of ρ, i.e.

$$\rho^* = \mathrm{E}(\rho|\mathbf{y}) \tag{4.43}$$

and all the expectations are taken with respect to the posterior distribution of (ξ, ϕ), or equivalently the posterior distribution of ρ.

There is a direct correspondence between ρ^* in (4.43) and the shrinkage parameter w^* in the Poisson/Gamma analysis of the preceding section. Here again, ρ^* gives the magnitude of shrinkage of the MLE towards the direction of the ratio $\frac{\mathrm{E}(\rho\xi|\mathbf{y})}{\mathrm{E}(\rho|\mathbf{y})}$.

The next step is to obtain the posterior density of (ξ, ϕ). We will consider the parametrisation in (3.2), that is $\xi = e^{\mu+\frac{1}{2}\sigma^2}$, $\phi = \xi^2\left(e^{\sigma^2}-1\right)$. In doing so we need only alter model (4.1a) by assigning a hyperprior distribution to the parameters (ξ, ϕ) rather than to (μ, σ^2). We take ξ and ϕ to be independently distributed according to vague uniform priors over $(0, \infty)$, i.e. $\pi(\xi, \phi) \propto 1$. Then their posterior distribution will be given by

$$\begin{aligned} p(\xi, \phi|\mathbf{y}) &\propto f(\mathbf{y}|\xi, \phi)\ \pi(\xi, \phi) \\ &\propto f(\mathbf{y}|\xi, \phi) \end{aligned}$$

where $f(\mathbf{y}|\xi, \phi)$ denotes the marginal density of the data given the hyperparameters ξ and ϕ. Obtaining this marginal density requires integrating $\boldsymbol{\theta}$ out of the joint conditional density of $(\mathbf{y}, \boldsymbol{\theta})$. According to the model, the latter is

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\theta}|\xi, \phi) &= f(\mathbf{y}|\boldsymbol{\theta})\ \pi(\boldsymbol{\theta}|\xi, \phi) \\ &\propto \prod_{i=1}^{m}\left[\frac{\theta_i^{y_i}e^{-\theta_i}}{y_i!}\ \frac{1}{\theta_i}\exp\left\{-\frac{1}{2}\sigma^{-2}(\log\theta_i - \mu)^2\right\}\right], \end{aligned}$$

defined for $\mathbf{y}$ and $\boldsymbol{\theta}$ being m-dimensional vectors of nonnegative integers and positive reals respectively. The above form implies an intractable integral with respect to $\boldsymbol{\theta}$, meaning that the marginal density of the data, and hence the posterior of the hyperparameters (ξ, ϕ), can not be derived in closed analytical form.

4.3.2 Normal approximation to the marginal data distribution

To tackle this problem we will assume a normal approximation to the marginal distribution of $\mathbf{y}$, as we did earlier with the conjugate model. The exact first two marginal moments of Y_i are derived in (3.20) and (3.21) as $\mathrm{E}(Y_i) = \xi$ and $\mathrm{var}(Y_i) = \xi + \phi$. If we let $\lambda = \xi + \phi$, the normal approximation is

$$Y_i \overset{iid}{\sim} N(\xi, \lambda), \qquad i = 1, \ldots, m, \tag{4.44}$$

and therefore the marginal density of $\mathbf{y}$ is approximately given by

$$\begin{aligned} f(\mathbf{y}|\xi, \lambda) &\propto \prod_{i=1}^{m}\left[\lambda^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\lambda^{-1}(y_i - \xi)^2\right\}\right] \\ &\propto \lambda^{-\frac{m}{2}}\exp\left\{-\frac{1}{2}\lambda^{-1}\sum_{i=1}^{m}(y_i - \xi)^2\right\}, \end{aligned} \tag{4.45}$$

for $\mathbf{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$ where y_i are nonnegative integers, and $0 < \xi < \lambda < \infty$. The prior distribution on (ξ, ϕ) is invariant to the reparametrisation (ξ, λ), and consequently the hyperprior density of (ξ, λ) is

$$\pi(\xi, \lambda) \propto 1. \tag{4.46}$$

Therefore, from Bayes' theorem and using (4.45) and (4.46) we have that the joint posterior distribution of the hyperparameters (ξ, λ) is approximately

$$\begin{aligned} p(\xi, \lambda|\mathbf{y}) &\propto f(\mathbf{y}|\xi, \lambda)\, \pi(\xi, \lambda) \\ &\propto \lambda^{-\frac{m}{2}} \exp\left\{-\frac{1}{2}\lambda^{-1}\sum_{i=1}^{m}(y_i - \xi)^2\right\}, \end{aligned} \tag{4.47}$$

which, by decomposing $(y_i - \xi)$ to $\{(y_i - \bar{y}) - (\xi - \bar{y})\}$, can be also written as

$$p(\xi, \lambda|\mathbf{y}) \propto \lambda^{-\frac{m}{2}} \exp\left\{-\frac{1}{2}\lambda^{-1}S_{yy} - \frac{1}{2}\lambda^{-1}m(\xi - \bar{y})^2\right\}, \tag{4.48}$$

where S_{yy} is the sum of squares corrected for the sample mean. We will now use the joint posterior density of ξ and λ given in (4.47) to derive the shrinkage proportion ρ^* appearing in the approximate posterior mean (4.42) of θ_i. This can be expressed in terms of ξ and λ as

$$\begin{aligned} \rho^* &= \mathrm{E}(\rho|\mathbf{y}) = \mathrm{E}\left(\frac{\xi}{\xi + \phi} \,\middle|\, \mathbf{y}\right) \\ &= \mathrm{E}\left(\frac{\xi}{\lambda} \,\middle|\, \mathbf{y}\right). \end{aligned}$$

We first attempt to obtain the posterior density of $\frac{\xi}{\lambda}$ using the bivariate transformation

$$\begin{aligned} \rho &= h(\xi, \lambda) = \frac{\xi}{\lambda}, \qquad 0 < \rho < 1 \\ \lambda &= g(\xi, \lambda) = \lambda, \qquad \xi < \lambda < \infty. \end{aligned} \tag{4.49}$$

Then, by the theory of transformations of multivariate random variables, we have that

$$p_{\rho,\lambda}(\rho, \lambda|\mathbf{y}) = p_{\xi,\lambda}(\xi, \lambda|\mathbf{y})\ |J|$$

where J denotes the Jacobian determinant, given by

$$J = det\begin{pmatrix} \frac{\partial \xi}{\partial \rho} & \frac{\partial \xi}{\partial \lambda} \\ \frac{\partial \lambda}{\partial \rho} & \frac{\partial \lambda}{\partial \lambda} \end{pmatrix} = det\begin{pmatrix} \lambda & \rho \\ 0 & 1 \end{pmatrix} = \lambda.$$

Hence the joint posterior density of (ρ, λ) is approximately given as

$$p(\rho, \lambda|\mathbf{y}) \propto \lambda^{-\frac{m}{2}+1} \exp\left\{-\frac{1}{2}\lambda^{-1}\sum_{i=1}^{m}(y_i - \rho\lambda)^2\right\}, \tag{4.50}$$

$0 < \rho < 1,\ \xi < \lambda < \infty$. The derivation of the posterior expectation of $\rho = \frac{\xi}{\lambda}$ requires to integrate λ out of the joint posterior density of (ρ, λ) and then average with respect to the marginal posterior of ρ, i.e

$$\rho^* = \mathrm{E}\left(\frac{\xi}{\lambda} \mid \mathbf{y}\right) = \int_0^1 \int_\xi^\infty \rho\, p(\rho, \lambda|\mathbf{y})\, d\lambda d\rho \tag{4.51}$$

which is a complicated task due to the form of $p(\rho, \lambda|\mathbf{y})$ in (4.50).

4.3.3 Approximate $\mathrm{E}(\theta_i|\mathbf{y})$ based on conditional expectations

An alternative way to derive ρ^* is to obtain the posterior expectation of $\rho = \frac{\xi}{\lambda}$ conditional on the hyperparameter λ and then average the conditional mean with respect to marginal posterior distribution of λ, that is

$$\rho^* = \mathrm{E}_\lambda\left\{\mathrm{E}(\rho|\lambda, \mathbf{y})\right\}. \tag{4.52}$$

We first calculate the conditional posterior density of ρ given λ from the joint posterior $p(\rho, \lambda|\mathbf{y})$. Notice that working as for (4.48), the joint posterior in (4.50) can be written as

$$p(\rho, \lambda|\mathbf{y}) \propto \lambda^{-\frac{m}{2}+1} \exp\left(-\frac{1}{2}\lambda^{-1}S_{yy}\right) \exp\left\{-\frac{1}{2}\lambda m\left(\rho - \frac{\bar{y}}{\lambda}\right)^2\right\},$$

$0 < \rho < 1,\ \xi < \lambda < \infty$, where S_{yy} is again the corrected sum of squares. If we ignore the terms not involving ρ, it follows that

$$p(\rho|\lambda, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\lambda m\left(\rho - \frac{\bar{y}}{\lambda}\right)^2\right\},$$

showing that the approximate posterior distribution of ρ given λ is normal with mean $\frac{\bar{y}}{\lambda}$ and variance $(\lambda m)^{-1}$, truncated to the interval $(0, 1)$, that is

$$\rho|\lambda, \mathbf{y} \sim N\left(\frac{\bar{y}}{\lambda}, \frac{1}{\lambda m}\right), \qquad 0 < \rho < 1. \tag{4.53}$$

We now give the following lemma regarding the moments of a truncated normal distribution.

Lemma 4.2. *Suppose that X is a truncated normal random variable. More specifically, we allow $X \sim N(\mu, \sigma^2)$ subject to the constraint $a \leq X \leq b$. Let $f(\cdot)$, $F(\cdot)$ denote the probability density function (p.d.f.) and the cumulative density function (c.d.f.) of X respectively, and $f_o(\cdot)$, $F_o(\cdot)$ the p.d.f. and the c.d.f. of an unrestricted normal variable. Then the following results hold (e.g. see Johnson, Kotz and Balakrishnan, 1994).*
The c.d.f. of X is given by

$$F(x) = \begin{cases} 0, & \text{if } x < a; \\ \frac{F_o(x)-F_o(a)}{F_o(b)-F_o(a)}, & \text{if } a \leq X \leq b; \\ 1, & \text{if } x > b \end{cases}$$

and thus, its p.d.f. is

$$f(x) = \begin{cases} \frac{f_o(x)}{F_o(b)-F_o(a)}, & \text{for } a \leq X \leq b; \\ 0, & \text{otherwise.} \end{cases}$$

Also, the first moment of X is given by

$$E(X) = \mu - \sigma^2 \frac{f_o(b) - f_o(a)}{F_o(b) - F_o(a)}, \qquad a \leq X \leq b, \tag{4.54}$$

and if we let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the p.d.f. and c.d.f. of a standardised normal variable respectively, the above can be written as

$$E(X) = \mu - \sigma \frac{\phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}. \tag{4.55}$$

Finally, the second moment of X will be

$$E(X^2) = \mu^2 + \sigma^2 \left\{ 1 - \frac{(b+\mu) f_o(b) - (a+\mu) f_o(a)}{F_o(b) - F_o(a)} \right\}, \tag{4.56}$$

or in terms of the standardised normal distribution

$$E(X^2) = \mu^2 + \sigma^2 \left\{ 1 - \frac{\left(\frac{b+\mu}{\sigma}\right) \phi\left(\frac{b-\mu}{\sigma}\right) - \left(\frac{a+\mu}{\sigma}\right) \phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \right\}. \tag{4.57}$$

Using the above results we can now derive the approximate conditional posterior expectations of $\frac{\xi}{\lambda}$ and $\frac{\xi^2}{\lambda}$. According to the approximation (4.53), the former is equal to the first moment of a normal $N\left(\frac{\bar{y}}{\lambda}, \frac{1}{\lambda m}\right)$ distribution truncated to the interval $(0, 1)$, and hence is given from (4.55) of Lemma 4.2 as

$$\mathrm{E}(\rho|\lambda, \mathbf{y}) = \mathrm{E}\left(\frac{\xi}{\lambda}\middle|\lambda, \mathbf{y}\right) \doteq \lambda^{-1} \left\{ \bar{y} - \sqrt{\frac{\lambda}{m}} \frac{\phi\left(\frac{\lambda-\bar{y}}{\sqrt{\lambda/m}}\right) - \phi\left(\frac{\bar{y}}{\sqrt{\lambda/m}}\right)}{\Phi\left(\frac{\lambda-\bar{y}}{\sqrt{\lambda/m}}\right) - \Phi\left(-\frac{\bar{y}}{\sqrt{\lambda/m}}\right)} \right\}. \tag{4.58}$$

We also notice that $\mathrm{E}\left(\frac{\xi^2}{\lambda}|\lambda,\mathbf{y}\right)$ is equal to $\lambda\ \mathrm{E}\left(\frac{\xi^2}{\lambda^2}|\lambda,\mathbf{y}\right)$ and therefore is given by the second moment of the truncated normal $N\left(\frac{\bar{y}}{\lambda},\frac{1}{\lambda m}\right)$ distribution multiplied by λ. Thus, from (4.57) we have that

$$\begin{aligned}\mathrm{E}(\rho\xi|\lambda,\mathbf{y}) &= \mathrm{E}\left(\frac{\xi^2}{\lambda}\Big|\lambda,\mathbf{y}\right)\\ &\doteq \lambda\left[\frac{\bar{y}^2}{\lambda^2}+\frac{1}{\lambda m}\left\{1-\frac{\left(\frac{\lambda+\bar{y}}{\sqrt{\lambda/m}}\right)\phi\left(\frac{\lambda-\bar{y}}{\sqrt{\lambda/m}}\right)-\left(\frac{\bar{y}}{\sqrt{\lambda/m}}\right)\phi\left(\frac{\bar{y}}{\sqrt{\lambda/m}}\right)}{\Phi\left(\frac{\lambda-\bar{y}}{\sqrt{\lambda/m}}\right)-\Phi\left(-\frac{\bar{y}}{\sqrt{\lambda/m}}\right)}\right\}\right]\\ &= \lambda^{-1}\left[\bar{y}^2+\frac{\lambda}{m}\left\{1-\frac{\left(\frac{\lambda+\bar{y}}{\sqrt{\lambda/m}}\right)\phi\left(\frac{\lambda-\bar{y}}{\sqrt{\lambda/m}}\right)-\left(\frac{\bar{y}}{\sqrt{\lambda/m}}\right)\phi\left(\frac{\bar{y}}{\sqrt{\lambda/m}}\right)}{\Phi\left(\frac{\lambda-\bar{y}}{\sqrt{\lambda/m}}\right)-\Phi\left(-\frac{\bar{y}}{\sqrt{\lambda/m}}\right)}\right\}\right] \quad (4.59)\end{aligned}$$

Remember that, as indicated in (4.42), the ratio $\frac{\mathrm{E}(\rho\xi|\mathbf{y})}{\mathrm{E}(\rho|\mathbf{y})}$ provides the point towards which our estimate $\widetilde{\mathrm{E}}(\theta_i|\mathbf{y})$ shrinks the MLE. Our approximate analysis, leading to the conditional expectations (4.58) and (4.59), suggests that this point approaches the data sample mean as m, the number of Poisson parameters, tends to infinity. However, it can be different than $\bar{y}$ when m is small. This is in agreement with our conclusion about the shrinkage behaviour of the EB posterior mean in Chapter 3.

To obtain the unconditional approximate posterior means of ρ and $\rho\xi$ we must average the conditional expectations in (4.58) and (4.59) with respect to the posterior distribution of the hyperparameter λ. We first derive this posterior using the approximate joint posterior distribution of (ξ,λ) given in (4.48). We have

$$p(\lambda|\mathbf{y}) = \int_0^\lambda p(\xi,\lambda|\mathbf{y})\ d\xi$$

and approximately

$$p(\lambda|\mathbf{y}) \propto \int_0^\lambda \lambda^{-\frac{m}{2}}\exp\left\{-\frac{1}{2}\lambda^{-1}S_{yy}-\frac{1}{2}\lambda^{-1}m(\xi-\bar{y})^2\right\}\ d\xi$$

$$\propto \lambda^{-\frac{m}{2}}\exp\left\{-\frac{1}{2}\lambda^{-1}S_{yy}\right\}\ \int_0^\lambda \exp\left\{-\frac{1}{2}\lambda^{-1}m(\xi-\bar{y})^2\right\}\ d\xi.$$

The integral in the last expression can be written in terms of the cdf of a normal $N(\bar{y},\frac{\lambda}{m})$ variate, denoted as $F\left(\cdot\ ;\bar{y},\frac{\lambda}{m}\right)$, and therefore the posterior density of λ

is approximately

$$p(\lambda|\mathbf{y}) \propto \lambda^{-\frac{m}{2}} \exp\left\{-\frac{1}{2}\lambda^{-1}S_{yy}\right\} \left(\frac{\lambda}{m}\right)^{\frac{1}{2}} \left\{F\left(\lambda; \bar{y}, \frac{\lambda}{m}\right) - F\left(0; \bar{y}, \frac{\lambda}{m}\right)\right\}$$

$$\propto \lambda^{-\frac{1}{2}(m-1)} \exp\left\{-\frac{1}{2}\lambda^{-1}S_{yy}\right\} \left\{F\left(\lambda; \bar{y}, \frac{\lambda}{m}\right) - F\left(0; \bar{y}, \frac{\lambda}{m}\right)\right\}$$

or, in terms of the cdf of a standardised normal variable

$$p(\lambda|\mathbf{y}) \propto \lambda^{-\frac{1}{2}(m-1)} \exp\left\{-\frac{1}{2}\lambda^{-1}S_{yy}\right\} \left\{\Phi\left(\frac{\lambda - \bar{y}}{\sqrt{\lambda/m}}\right) - \Phi\left(-\frac{\bar{y}}{\sqrt{\lambda/m}}\right)\right\}, \quad (4.60)$$

where $\xi < \lambda < \infty$.

We can finally obtain our approximation to the posterior expectation of the Poisson means $\theta_1, \theta_2, \ldots, \theta_m$, using the linear estimator (4.41), where the unconditional posterior expectations of ρ and $\rho\xi$ are calculated by appropriate averagings as indicated in (4.52), with $\mathrm{E}(\rho|\lambda, \mathbf{y})$,$\mathrm{E}(\rho\xi|\lambda, \mathbf{y})$ and $p(\lambda|\mathbf{y})$ given in equations (4.58), (4.59) and (4.60) respectively.

4.4 Summary and conclusions

In this chapter we have considered a hierarchical Bayes approach for the analysis of Poisson models. Two hierarchical formulations were considered, one being the Poisson/Gamma first stage conjugate structure, and the other the Poisson/log-normal model.

The hierarchical nature of both settings implies that analytically explicit exact solutions cannot be obtained. We have therefore attempted to derive a full Bayesian solution employing analytical approximations. To obtain the unconditional posterior means of the Poisson parameters θ_i, $i = 1, \ldots, m$, in the Poisson/gamma model, we consider a normal approximation to the marginal distribution of the data, using the first two moments of the exact negative binomial marginal distribution. A further approximation based on the χ^2 statistic is employed in order to obtain a simpler expression for the posterior mean. The implementation of the approximate method requires the use of numerical integration, at the final stage of its evaluation. The approximate posterior mean suggests that the hierarchical Bayes analysis results in an estimator that shrinks less than the EB methods, in the case where the estimated parameters are close together.

For the Poisson/log-normal model we have assumed vague prior distributions for the log-normal hyperparameters, and a linear approximation to the posterior mean conditional on these hyperparameters is considered, following the derivation

of the BLP for the Poisson/log-normal model in Chapter 3. Furthermore, a normal approximation to the marginal nonclosed distribution of the data is assumed, employing again the exact mean and variance. The unconditional posterior means of the parameters θ_i, $i = 1, \ldots, m$, are still given in mathematically intractable form. A more simplified form can be reached, considering the results conditioned on the normal variance, but again substantial numerical work is needed. However, the form of the conditional estimator of θ_i, $i = 1, \ldots, m$, suggests that, for small or moderate m, the approximate posterior mean does not shrink the observed values towards the sample mean.

It is important to stress here that the approximate analysis and results for the Poisson/log-normal model in this chapter rely on two approximations: that of the linear estimator to the posterior mean and a normal approximation for the marginal distribution of the data. Moreover, the computation of the resulting estimators rests on several implementations of numerical integration techniques. An alternative is to use Monte Carlo integration methods for the estimation of the Poisson means in our hierarchical model. These are the subject of the following chapters of this thesis.

Chapter 5

Simulation methods for the full hierarchical Bayesian analysis

5.1 Introduction

As emphasised in Chapter 4, the full hierarchical analysis of the Poisson/log-normal model, can only be approximately performed in analytical form. Aiming at the exact analysis, we may alternatively attempt to tackle the inferential problem using Monte Carlo integration techniques. The basis of the Monte Carlo integration methodology is to obtain a sufficient number of simulations from the distribution of interest and then estimate the required characteristics of that distribution using the generated values. In the Bayesian context the distribution under consideration will be the posterior distribution of the parameters of interest, and therefore we want to use the simulated values to obtain the characteristics of the posterior distribution, such as moments, quantiles, confidence intervals, marginal and predictive densities etc. Simulating directly from the posterior distribution might be difficult, or even impossible when the distribution of interest does not appear in a closed analytical form. Several Monte Carlo methods have been proposed to deal with such cases, including importance sampling and Markov chain Monte Carlo, which we will discuss in the present chapter.

5.2 Importance sampling

In Chapter 3 we employed the importance sampling technique to estimate the conditional expectations of the Poisson means $\mathrm{E}(\theta_i|\mu,\sigma^2,\mathbf{y})$, $i=1,\ldots,m$. The idea was to express the conditional posterior density of $\gamma_i=\log(\theta_i)$ as

$$p(\gamma_i|\mu,\sigma^2,\mathbf{y}) \propto I(\gamma_i)\ W(\gamma_i), \qquad i=1,\ldots,m,$$

where

$$I(\gamma_i) = \exp\{\gamma_i(y_i + 1) - e^{\gamma_i}\},$$

and

$$W(\gamma_i) = \exp\left\{-\frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2 - \gamma_i\right\}.$$

Notice that the quantity $I(\gamma_i)$ is proportional to the density of γ_i, where e^{γ_i} follows a $\text{Ga}(y_i + 1, 1)$ distribution. We therefore approximated the conditional posterior expectations by averaging over the generated values $e^{\gamma_{ij}} W(\gamma_{ij})$, $j = 1, \ldots, N$, where the γ_{ij} variates were simulated such that $e^{\gamma_{ij}}$, $j = 1, \ldots, N$, independently follow a $\text{Ga}(y_i + 1, 1)$ distribution.

We will investigate whether we can apply a similar approach to the estimation of the unconditional posterior means in the full hierarchical Poisson/log-normal model. We will consider both prior specifications for the hyperparameter σ^2, that is the flat uniform prior distribution over the interval $(0, \infty)$ and the Inv-$\chi^2(\nu, \lambda)$ prior distribution.

5.2.1 Model (4.1a): Uniform hyperprior on σ^2

Under model (4.1a) we can write the joint posterior density of the parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_m)^{\mathrm{T}}$, μ and σ^2, as

$$\begin{aligned} p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\gamma}, \mu, \sigma^2)\ \pi(\boldsymbol{\gamma}|\mu, \sigma^2)\ \pi(\mu, \sigma^2) \\ &\propto (\sigma^2)^{-\frac{1}{2}m} \exp\left[\sum_{i=1}^{m}\left\{\gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2\right\}\right], \end{aligned}$$

where $-\infty < \gamma_i < \infty$, for $i = 1, \ldots, m$, $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$. Now, if we decompose $(\gamma_i - \mu)$ to $(\gamma_i - \bar{\gamma}) - (\mu - \bar{\gamma})$, and after some rearrangement, the above can be written as

$$\begin{aligned} p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}m} \exp\left\{\sum_{i=1}^{m}\gamma_i(y_i + 1) - \sum_{i=1}^{m} e^{\gamma_i}\right\}\ e^{-\sum_{i=1}^{m}\gamma_i} \\ \exp\left\{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}\ \exp\left\{-\frac{1}{2}\sigma^{-2}m\ (\mu - \bar{\gamma})^2\right\}. \quad (5.1) \end{aligned}$$

We must integrate both μ and σ^2 out of the joint posterior distribution (5.1) in order to obtain the unconditional posterior density of the vector parameter $\boldsymbol{\gamma}$. We first derive $p(\boldsymbol{\gamma}, \sigma^2|\mathbf{y})$ as following:

$$p(\boldsymbol{\gamma}, \sigma^2|\mathbf{y}) = \int_{-\infty}^{\infty} p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y})\ d\mu,$$

and substituting (5.1) in the above we obtain

$$p(\boldsymbol{\gamma}, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}m} \exp\left\{\sum_{i=1}^{m} \gamma_i(y_i+1) - \sum_{i=1}^{m} e^{\gamma_i}\right\} e^{-\sum_{i=1}^{m}\gamma_i}$$
$$\exp\left\{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2\right\} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\sigma^{-2}m\,(\mu-\bar{\gamma})^2\right\} d\mu.$$

The integrand involved in the above expression is proportional to a normal $N(\bar{\gamma}, \frac{\sigma^2}{m})$ density, where $\bar{\gamma} = m^{-1}\sum_{i=1}^{m}\gamma_i$, and therefore the joint posterior density of $\boldsymbol{\gamma}$ and σ^2 is given by

$$p(\boldsymbol{\gamma}, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}(m-1)} \exp\left\{\sum_{i=1}^{m} \gamma_i(y_i+1) - \sum_{i=1}^{m} e^{\gamma_i}\right\} e^{-\sum_{i=1}^{m}\gamma_i}$$
$$\exp\left\{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2\right\}$$
$$= I(\boldsymbol{\gamma})\, e^{-\sum_{i=1}^{m}\gamma_i}\, (\sigma^2)^{-\frac{1}{2}(m-1)} \exp\left\{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2\right\}, \qquad (5.2)$$

where

$$I(\boldsymbol{\gamma}) = \exp\left\{\sum_{i=1}^{m} \gamma_i(y_i+1) - \sum_{i=1}^{m} e^{\gamma_i}\right\}. \qquad (5.3)$$

We now obtain the marginal posterior density of $\boldsymbol{\gamma}$ by integrating σ^2 out of (5.2). This will give

$$p(\boldsymbol{\gamma}|\mathbf{y}) = \int_0^{\infty} p(\boldsymbol{\gamma}, \sigma^2|\mathbf{y})\, d\sigma^2$$
$$\propto I(\boldsymbol{\gamma})\, e^{-\sum_{i=1}^{m}\gamma_i} \int_0^{\infty} (\sigma^2)^{-\frac{1}{2}(m-1)} \exp\left\{-\frac{1}{2}\sigma^{-2}\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2\right\} d\sigma^2.$$

Using the variable transformation $\tau^2 = \left\{\frac{\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2}{\sigma^2}\right\}^{-1}$, we have that

$$p(\boldsymbol{\gamma}|\mathbf{y}) = I(\boldsymbol{\gamma})\, e^{-\sum_{i=1}^{m}\gamma_i} \left\{\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2\right\}^{-\frac{1}{2}(m-3)}$$
$$\int_0^{\infty} (\tau^2)^{-(\frac{m-3}{2}+1)} \exp(-\frac{1}{2}\tau^{-2})\, d\tau^2.$$

We notice that the integrand in the above expression is proportional to the probability density function of an inverse chi-square random variable τ^2, given in Appendix A, with $(m-3)$ degrees of freedom. Thus, the integral will be equal

to the reciprocal of a normalising constant not involving $\boldsymbol{\gamma}$, and therefore the marginal posterior density of $\boldsymbol{\gamma}$ is given by

$$p(\boldsymbol{\gamma}|\mathbf{y}) \propto I(\boldsymbol{\gamma})\, e^{-\sum_{i=1}^{m}\gamma_i} \left\{\sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}^{-\frac{1}{2}(m-3)} \tag{5.4}$$

$$= I(\boldsymbol{\gamma})\, W(\boldsymbol{\gamma}), \qquad \boldsymbol{\gamma} \in \mathbb{R}^m \tag{5.5}$$

where

$$W(\boldsymbol{\gamma}) = e^{-\sum_{i=1}^{m}\gamma_i} \left\{\sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}^{-\frac{1}{2}(m-3)}, \tag{5.6}$$

and $I(\boldsymbol{\gamma})$ is given in (5.3). Then, in principle, we can use the idea introduced in Chapter 3 to obtain the posterior expectation of a function of $\boldsymbol{\gamma}$. We first notice that the quantity $I(\boldsymbol{\gamma})$ is proportional to the joint probability density function of m independent random variables $\gamma_1, \gamma_2, \ldots, \gamma_m$, where for $i = 1, \ldots, m$, e^{γ_i} follows a $\mathrm{Ga}(y_i + 1, 1)$ distribution. Hence, the posterior expectation of $\theta_i = e^{\gamma_i}$ is given by

$$\begin{aligned}\mathrm{E}(\theta_i|\mathbf{y}) &= \int_{\mathbb{R}^m} e^{\gamma_i}\, p(\boldsymbol{\gamma}|\mathbf{y})\, d\boldsymbol{\gamma} \\ &= c^{-1} \int_{\mathbb{R}^m} e^{\gamma_i}\, W(\boldsymbol{\gamma})\, I(\boldsymbol{\gamma})\, d\boldsymbol{\gamma},\end{aligned}$$

with $c = \int_{\mathbb{R}^m} W(\boldsymbol{\gamma}) I(\boldsymbol{\gamma}) d\boldsymbol{\gamma}$ being the normalising factor. It follows that

$$\mathrm{E}(\theta_i|\mathbf{y}) = \frac{\int_{\mathbb{R}^m} e^{\gamma_i}\, W(\boldsymbol{\gamma})\, I(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}}{\int_{\mathbb{R}^m} W(\boldsymbol{\gamma})\, I(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}}$$

$$= \frac{\mathrm{E}_I\{e^{\gamma_i} W(\boldsymbol{\gamma})\}}{\mathrm{E}_I\{W(\boldsymbol{\gamma})\}}. \tag{5.7}$$

Here, E_I denotes the expectation with respect to the distribution whose probability density function is given by $I(\boldsymbol{\gamma})$ multiplied by the appropriate normalising constant. We will refer to this normalised density as the importance density, and to the function $W(\boldsymbol{\gamma})$ as the weight function. Then, under the conditions stated in Section 3.4, as $N \to \infty$, expression (5.7) leads to the following importance sampling approximation to the posterior mean of θ_i:

$$\mathrm{E}^{\mathrm{IS}}(\theta_i|\mathbf{y}) = \frac{\sum_{j=1}^{N} e^{\gamma_{ij}} W(\boldsymbol{\gamma}_j)}{\sum_{j=1}^{N} W(\boldsymbol{\gamma}_j)}, \tag{5.8}$$

where

$$W(\boldsymbol{\gamma}_j) = e^{-\sum_{i=1}^{m}\gamma_{ij}} \left\{\sum_{i=1}^{m}(\gamma_{ij} - \bar{\gamma}_j)^2\right\}^{-\frac{1}{2}(m-3)}, \tag{5.9}$$

with $\bar{\gamma}_j = m^{-1}\sum_{i=1}^{m}\gamma_{ij}$, and $e^{\gamma_{ij}}$, $j = 1,\ldots,N$ being independent realisations from a $\mathrm{Ga}(y_i+1,1)$ distribution.

The estimator given in (5.8) is of practical use only when it can be computed with a finite standard error of simulation. Geweke (1989) shows that the standard error of simulation for the importance sampling estimator of $\mathrm{E}(\theta_i|\mathbf{y})$ can be estimated by

$$\frac{\left[\sum_{j=1}^{N}\left\{\theta_{ij}-\mathrm{E}^{\mathrm{IS}}(\theta_i|\mathbf{y})\right\}^2 W^2(\boldsymbol{\gamma}_j)\right]^{\frac{1}{2}}}{\sum_{j=1}^{N} W(\boldsymbol{\gamma}_j)}, \tag{5.10}$$

where $W(\boldsymbol{\gamma}_j)$ and the simulated values $\theta_{ij} = e^{\gamma_{ij}}$ are obtained as for (5.8). Geweke (1989) also shows that (5.10) will be finite when the posterior expectations of $W(\boldsymbol{\gamma})$ and $(e^{\gamma_i})^2 W(\boldsymbol{\gamma})$ are finite, that is when

$$\mathrm{E}\{W(\boldsymbol{\gamma})|\mathbf{y}\} = \int_{\mathbb{R}^m} W(\boldsymbol{\gamma})\; p(\boldsymbol{\gamma}|\mathbf{y})\; d\boldsymbol{\gamma} < \infty \tag{5.11}$$

and

$$\mathrm{E}\left\{(e^{\gamma_i})^2 W(\boldsymbol{\gamma})|\mathbf{y}\right\} = \int_{\mathbb{R}^m} e^{2\gamma_i}\; W(\boldsymbol{\gamma})\; p(\boldsymbol{\gamma}|\mathbf{y})\; d\boldsymbol{\gamma} < \infty. \tag{5.12}$$

The above conditions are satisfied when the corresponding integrands are bounded. Thus, using (5.5), we require the functions

$$W^2(\boldsymbol{\gamma})\; I(\boldsymbol{\gamma}) = \left[e^{-\sum_{i=1}^{m}\gamma_i}\left\{\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2\right\}^{-\frac{1}{2}(m-3)}\right]^2 \exp\left\{\sum_{i=1}^{m}\gamma_i(y_i+1)-\sum_{i=1}^{m}e^{\gamma_i}\right\} \tag{5.13}$$

and

$$\{e^{\gamma_i}W(\boldsymbol{\gamma})\}^2\; I(\boldsymbol{\gamma}) = \left[e^{\gamma_i}e^{-\sum_{i=1}^{m}\gamma_i}\left\{\sum_{i=1}^{m}(\gamma_i-\bar{\gamma})^2\right\}^{-\frac{1}{2}(m-3)}\right]^2 \exp\left\{\sum_{i=1}^{m}\gamma_i(y_i+1)-\sum_{i=1}^{m}e^{\gamma_i}\right\} \tag{5.14}$$

to be bounded. Unfortunately, this is not always true, as the functions (5.13) and (5.14) may tend to infinity for relatively large negative values of the variates γ_i. Since the γ_i values are generated in such a manner that $e^{\gamma_i} \sim \mathrm{Ga}(y_i+1,1)$, large negative values of γ_i are likely to occur when small observations y_i are involved, implying that the estimator (5.8) will probably behave poorly for data

sets containing many zero observations. The behaviour of functions (5.13) and (5.14) may be improved when the sample size m is large, although any possible improvement will still be reduced or even eliminated by the presence of many zero values in the data.

5.2.2 Model (4.1b): Inv-$\chi^2(\nu, \lambda)$ hyperprior on σ^2

We will now consider model (4.1b). In this case we assume that the hyperparameter σ^2 has a scaled inverse chi-square prior distribution with parameters ν and λ and therefore, the joint posterior density of the vector parameter $\boldsymbol{\gamma}$ and the hyperparameters μ and σ^2 is now written as

$$\begin{aligned} p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\gamma}, \mu, \sigma^2)\ \pi(\boldsymbol{\gamma}|\mu, \sigma^2)\ \pi(\mu, \sigma^2) \\ &\propto (\sigma^2)^{-\frac{1}{2}m} \exp\left[\sum_{i=1}^{m}\left\{\gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2\right\}\right] \\ &\quad (\sigma^2)^{-(\frac{\nu}{2}+1)} \exp\left(-\frac{1}{2}\frac{\nu\lambda}{\sigma^2}\right), \end{aligned}$$

where $-\infty < \gamma_i < \infty$, $i = 1, \ldots, m$, $-\infty < \mu < \infty$ and $0 < \sigma^2 < \infty$. Working as for (5.1) the above expression becomes

$$\begin{aligned} p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y}) &\propto (\sigma^2)^{-(\frac{\nu+m}{2}+1)}\ I(\boldsymbol{\gamma})\ e^{-\sum_{i=1}^{m}\gamma_i} \\ &\exp\left[-\frac{1}{2}\sigma^{-2}\left\{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}\right]\ \exp\left\{-\frac{1}{2}\sigma^{-2}\ m\ (\mu - \bar{\gamma})^2\right\}, \end{aligned} \tag{5.15}$$

where $I(\boldsymbol{\gamma})$ is given in (5.3). To derive the marginal posterior density of $\boldsymbol{\gamma}$ we integrate (5.15) over μ and σ^2 successively. The first integration is similar to that leading to (5.2), and hence it will give

$$\begin{aligned} p(\boldsymbol{\gamma}, \sigma^2|\mathbf{y}) &= \int_{-\infty}^{\infty} p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y})\ d\mu \\ &\propto I(\boldsymbol{\gamma})\ e^{-\sum_{i=1}^{m}\gamma_i}\ (\sigma^2)^{(-\frac{\nu+m+1}{2})} \\ &\quad \exp\left[-\frac{1}{2}\sigma^{-2}\left\{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}\right]. \end{aligned} \tag{5.16}$$

Then, integrating with respect to σ^2 we obtain

$$\begin{aligned} p(\boldsymbol{\gamma}|\mathbf{y}) &= \int_0^{\infty} p(\boldsymbol{\gamma}, \sigma^2|\mathbf{y})\ d\sigma^2 \\ &\propto I(\boldsymbol{\gamma})\ e^{-\sum_{i=1}^{m}\gamma_i} \int_0^{\infty} (\sigma^2)^{(-\frac{\nu+m+1}{2})} \exp\left[-\frac{1}{2}\sigma^{-2}\left\{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}\right]\ d\sigma^2. \end{aligned}$$

If we now apply the variable transformation $u^2 = \left\{\frac{\nu\lambda+\sum_{i=1}^m(\gamma_i-\bar{\gamma})^2}{\sigma^2}\right\}^{-1}$, the above is written as

$$p(\boldsymbol{\gamma}|\mathbf{y}) = I(\boldsymbol{\gamma})\; e^{-\sum_{i=1}^m \gamma_i}\; \left\{\nu\lambda + \sum_{i=1}^m(\gamma_i-\bar{\gamma})^2\right\}^{-(\frac{\nu+m-1}{2})} \int_0^\infty (u^2)^{-(\frac{\nu+m-1}{2}+1)}\, \exp(-\frac{1}{2}u^{-2})\; du^2,$$

and since the integrand involved in the above expression is proportional to the probability density function of an inverse chi-square distribution with $(\nu+m-1)$ degrees of freedom, the marginal posterior density of $\boldsymbol{\gamma}$ is given by

$$p(\boldsymbol{\gamma}|\mathbf{y}) \propto I(\boldsymbol{\gamma})\; e^{-\sum_{i=1}^m \gamma_i}\; \left\{\nu\lambda + \sum_{i=1}^m(\gamma_i-\bar{\gamma})^2\right\}^{-(\frac{\nu+m-1}{2})}, \qquad \boldsymbol{\gamma}\in\mathbb{R}^m. \tag{5.17}$$

We notice that, as in Subsection 5.2.1, the posterior density of the vector parameter $\boldsymbol{\gamma}$ can be expressed as the product of the importance density $I(\boldsymbol{\gamma})$ and a new weight function $U(\boldsymbol{\gamma})$, i.e.

$$p(\boldsymbol{\gamma}|\mathbf{y}) \propto I(\boldsymbol{\gamma})\; U(\boldsymbol{\gamma}), \tag{5.18}$$

where $I(\boldsymbol{\gamma})$ is the same as in (5.3) and the weight function takes the form

$$U(\boldsymbol{\gamma}) = e^{-\sum_{i=1}^m \gamma_i}\; \left\{\nu\lambda + \sum_{i=1}^m(\gamma_i-\bar{\gamma})^2\right\}^{-(\frac{\nu+m-1}{2})}. \tag{5.19}$$

Then, in correspondence with the argument leading to estimator (5.8), the importance sampling approximation to the posterior mean of θ_i is now given as

$$\mathrm{E}^{\mathrm{IS}}(\theta_i|\mathbf{y}) = \frac{\sum_{j=1}^N e^{\gamma_{ij}} U(\boldsymbol{\gamma}_j)}{\sum_{j=1}^N U(\boldsymbol{\gamma}_j)}, \tag{5.20}$$

where

$$U(\boldsymbol{\gamma}_j) = e^{-\sum_{i=1}^m \gamma_{ij}}\; \left\{\nu\lambda + \sum_{i=1}^m(\gamma_{ij}-\bar{\gamma}_j)^2\right\}^{-(\frac{\nu+m-1}{2})}, \tag{5.21}$$

with γ_{ij}, $j = 1,\ldots,N$, simulated in such a way that the variates $e^{\gamma_{i1}},\ldots,e^{\gamma_{iN}}$, form an independent sample from a $\mathrm{Ga}(y_i+1,1)$ distribution.

The assumption of a scaled inverse chi-square prior distribution for the variance σ^2, rather than a uniform one in the previous subsection, has led to a different weight function for the importance sampling estimator. As a consequence, the required conditions for the estimates obtained with (5.20) to be computed with a

finite standard error of simulation, now imply that we need the posterior expectations of $U(\boldsymbol{\gamma})$ and $(e^{\gamma_i})^2 U(\boldsymbol{\gamma})$ to be finite. Thus, we require that the functions

$$U^2(\boldsymbol{\gamma})\ I(\boldsymbol{\gamma}) \qquad \text{and} \qquad \{e^{\gamma_i} U(\boldsymbol{\gamma})\}^2\ I(\boldsymbol{\gamma}) \tag{5.22}$$

are bounded. The form of $U(\boldsymbol{\gamma})$ in (5.19) implies that the functions (5.22) may be unbounded for extreme negative values of the components of vector $\boldsymbol{\gamma}$. However, for all practical purposes, we can choose the degrees of freedom ν, of the prior distribution of σ^2, in such a way that both functions (5.22) are bounded within the support of the importance density, that is for the entire range of values that $\boldsymbol{\gamma}$ can take in practice. Unfortunately though, when a large number of zeroes has been observed in the data, implying that $\boldsymbol{\gamma}$ contains large negative elements, the estimates resulting from (5.20) may exhibit considerable fluctuations even after a very large number of simulations as illustrated in the oilwell data example, analysed in a subsequent subsection.

5.2.3 Example: Audit data

We demonstrate the methods described in Subsections 5.2.1 and 5.2.2 using the two real data examples introduced in Chapter 3. The audit data set, presented in Section 3.6, concerns the number of errors found in audit samples of 9 different accounts. The observations, given in Table 3.1, are $0, 0, 0, 1, 1, 2, 2, 3, 6$. Here we will adopt a full hierarchical Bayesian approach and follow the methods presented in Subsections 5.2.1 and 5.2.2. According to the two models (4.1a) and (4.1b), we assume that each of the observations $Y_1, Y_2, \ldots, Y_9$, independently follows a Poisson distribution, given its respective mean $\theta_1, \theta_2, \ldots, \theta_m$. We set $\gamma_i = \log(\theta_i)$, and we assume that the variables $\gamma_1, \gamma_2, \ldots, \gamma_m$, are independently and normally $N(\mu, \sigma^2)$ distributed. Then, for the second stage of the prior setting, we consider two separate cases in correspondence to the two models (4.1a) and (4.1b). We first assume that μ and σ^2 are independent and they both follow vague uniform distributions over $(-\infty, \infty)$ and $(0, \infty)$ respectively. For the second case we assume that μ is again uniformly distributed, while σ^2 independently has a Inv-$\chi^2(\nu, \lambda)$ distribution. In the latter specification we assigned a crude informative prior to the prior variance σ^2, by taking $\nu = 10$ and $\lambda = 0.45$. The choice of these values was based on matching the mean of the scaled inverse chi-square distribution to a data driven estimate of σ^2, after fixing ν equal to 10.

The estimates from 3 implementations of the importance sampling algorithm, with $N = 2 \times 10^4$, 10^5 and 10^7 simulations respectively, are reported. We notice here that $N = 10^7$ will normally be a prohibitively high number of simulations, in terms of computer time cost.

Table 5.1: *Importance sampling estimates of the posterior mean for the audit data set when model (4.1a) is assumed. The standard error of simulation (5.10) is given in brackets.*

		$E^{IS}(\theta_i \mid \mathbf{y})$					
par.	y_i	$N = 2 \times 10^4$		$N = 10^5$		$N = 10^7$	
θ_1	0	0.665	(0.068)	0.515	(0.015)	0.477	(0.011)
θ_2	0	0.450	(0.114)	0.467	(0.056)	0.456	(0.022)
θ_3	0	0.504	(0.033)	0.511	(0.041)	0.483	(0.018)
θ_4	1	1.232	(0.039)	1.097	(0.026)	1.043	(0.026)
θ_5	1	1.272	(0.066)	1.139	(0.034)	1.071	(0.016)
θ_6	2	1.800	(0.091)	2.853	(0.652)	1.791	(0.028)
θ_7	2	1.762	(0.061)	1.761	(0.070)	1.793	(0.023)
θ_8	3	2.515	(0.082)	2.556	(0.071)	2.582	(0.029)
θ_9	6	5.116	(0.150)	4.986	(0.114)	5.178	(0.033)

The results obtained assuming a uniform prior for the variance parameter σ^2 and using the estimator (5.8), are presented in Table 5.1. The first block of columns contains the results with $N = 2 \times 10^4$ simulations, whereas the second and third blocks give the estimates when $N = 10^5$ and $N = 10^7$ simulations were involved respectively. The numbers in brackets provide the associated standard error of simulation, as the latter is computed using (5.10). Comparing the parameter estimates in the 3 blocks, it is obvious that even a number of $N = 10^5$ simulations is inadequate. The standard error of simulation is still remarkably high. The value 0.652 that the standard error takes for the estimate of θ_6, is indicative of a presumed unbounded weight function. Figure 5.1(a) displays the trace of the posterior mean of θ_1 as the simulation procedure progresses. The dotted line corresponds to the correct estimate of the posterior mean, as this is derived by the methods suggested in later sections. These estimates are $0.50, 1.09, 1.81, 2.59$ and 5.14 when y is equal to $0, 1, 2, 3$ and 6 respectively. Clearly, Figure 5.1(a) shows that the importance sampling estimator does not seem to provide an accurate result, even after $N = 10^6$ simulations. Furthermore, the sharp fluctuations appearing in the graph, suggest that the method is not reliable. The numbers in the last block of columns of Table 5.1 show that after $N = 10^7$ replications of the algorithm, the simulation standard error is still considerable, implying that

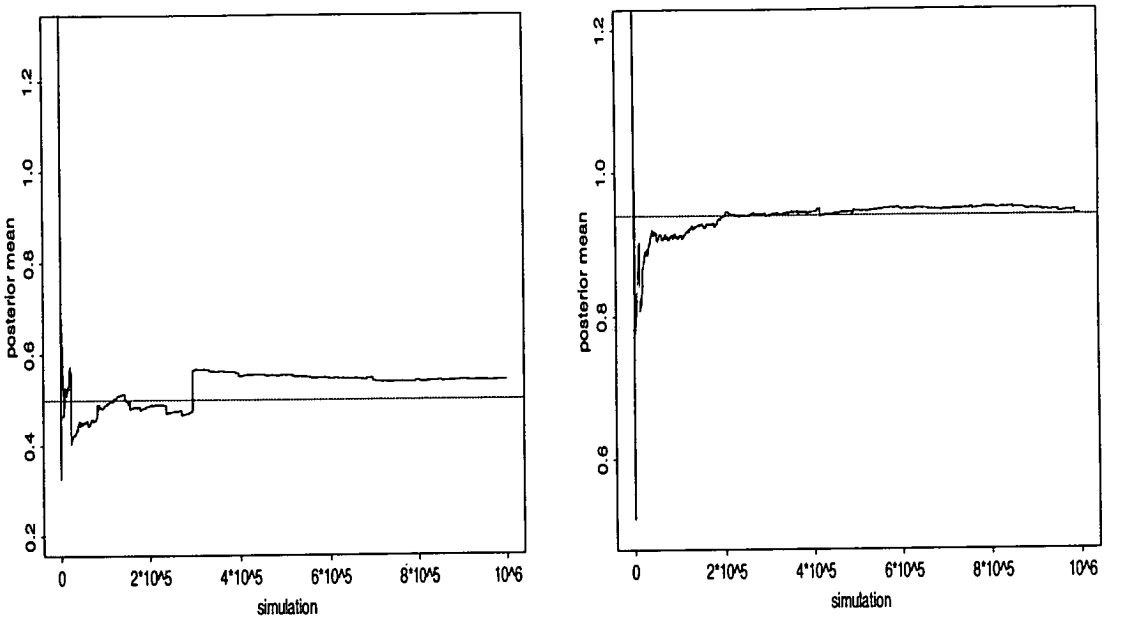

Figure 5.1: *Trace of* $E^{IS}(\theta_1|\mathbf{y})$*: (a) under model (4.1a); (b) under model (4.1b). The dotted line corresponds to the correct estimate of the posterior mean.*

the parameter estimates are not yet adequately close to the correct values. In fact, comparing the importance sampling results with the correct estimates given above, we notice that there still seems to exist some disagreement of a magnitude ranging from 0.01 to 0.05. This may suggest that, for this model, the importance sampling method that we have used fails to provide accurate estimates for the model parameters under any reasonable number of Monte Carlo simulations.

Table 5.2 contains the results of the analysis of the same data set, assuming an Inv-$\chi^2(10, 0.45)$ prior distribution for σ^2 and using the method described in Subsection 5.2.2. These results manifest that, under the light of stronger prior knowledge, the importance sampling estimator performs better, but still requires a remarkably high number of replications in order to provide the correct estimates. This is also illustrated in Figure 5.1(b), where the importance sampling estimate of the posterior mean of θ_1 seems to coincide with the correct value after around 3×10^5 simulations, but still exhibits some variation during the rest of the simulation process. Similar checks for the estimates of the other model parameters indicated that obtaining an accurate and reliable result might take more than 10^6 simulations. For this case the correct estimates are $0.94, 1.29, 1.71, 2.19$ and 3.97 for $y = 0, 1, 2, 3$ and 6 respectively. Clearly, the importance sampling estimator yields the correct estimates after 10^7 replications, as also indicated by

Table 5.2: *Importance sampling estimates of the posterior mean for the audit data set when model (4.1b) is assumed. The standard error of simulation (5.10) is given in brackets.*

		$E^{IS}(\theta_i\|\mathbf{y})$		
par.	y_i	$N = 2 \times 10^4$	$N = 10^5$	$N = 10^7$
θ_1	0	0.957 (0.048)	0.925 (0.035)	0.944 (0.003)
θ_2	0	1.032 (0.081)	0.976 (0.018)	0.938 (0.004)
θ_3	0	0.914 (0.037)	0.974 (0.015)	0.940 (0.004)
θ_4	1	1.230 (0.057)	1.316 (0.036)	1.296 (0.007)
θ_5	1	1.254 (0.093)	1.292 (0.026)	1.291 (0.006)
θ_6	2	1.818 (0.053)	1.822 (0.037)	1.719 (0.007)
θ_7	2	1.489 (0.205)	1.739 (0.030)	1.709 (0.009)
θ_8	3	2.169 (0.150)	2.108 (0.134)	2.193 (0.015)
θ_9	6	4.021 (0.170)	3.914 (0.182)	3.987 (0.012)

the low standard error of simulation in the last column of Table 5.2. However, as stressed before, the time cost for such an intensive computer simulation process is normally prohibitively high.

5.2.4 Example: Oilwell discoveries data

In this example the data, also considered in Section 3.7, record the number of oilwell discoveries in Canada for 36 months during the years 1953-1970. Two monthly readings, in March and September, are reported per year. The observed numbers of discoveries are displayed in Table 3.3. Clevenson and Zidek (1975) and Leonard (1976) propose two different shrinkage estimators for the estimation of the mean number of discoveries, while George, Makov and Smith (1994) consider a hierarchical Bayesian approach and a Markov chain Monte Carlo solution to the problem. Their approach differs from ours in that, rather than the conjugate Poisson/gamma formulation that they adopt, we consider the Poisson/log-normal modelling that we used for the audit data example. We also follow the same procedure as before, with the only alteration being that the hyperparameters for the Inv-$\chi^2(\nu, \lambda)$ prior distribution on σ^2 are now equal to $\nu = 10$ and $\lambda = 0.46$, determined in the same manner as in Subsection 5.2.3.

We attempted 3 separate implementations of the importance sampling algo-

Table 5.3: *Maximum value of importance sampling standard error of simulation for the oilwell discoveries data set after* 10^7 *Monte Carlo simulations.*

par.	y_i	*Model (4.1a)*	*Model (4.1b)*
θ_1-θ_{19}	0	0.201	0.196
θ_{20}-θ_{29}	1	0.161	0.356
θ_{30}-θ_{33}	2	0.351	0.227
θ_{34}, θ_{35}	3	0.545	0.228
θ_{36}	5	0.362	0.228

rithm, with 10^5, 10^6 and 10^7 simulations respectively. Unfortunately the outcome was disappointing, as even with a number of 10^7 simulations the method provided very bad results, with the posterior mean estimates exhibiting severe instability and the standard error of simulation taking unusually high values. Table 5.3 displays the maximum value of the standard error of simulation for the estimates of the Poisson means, after 10^7 simulations, and for both the distributional assumptions for the variance parameter σ^2 at the second stage of the prior specification. It is obvious that the method is not reliable, even when such a large number of replications is used. Furthermore, the assumption of an informative prior for σ^2 does not seem to offer considerable improvement, implying that the nature of this particular data set necessitates the use of a different approach. This suggests that the importance sampling technique that we presented in Subsections 5.2.1 and 5.2.2, based on the particular importance function that we proposed, cannot be regarded as a general-purpose method, and should not be used without taking into consideration the nature of each particular problem.

5.2.5 Importance sampling for marginal densities

Until now we have used the importance sampling method for the estimation of posterior means. We can employ a similar procedure to derive the marginal posterior density of any single component of the parameter vector $\boldsymbol{\gamma}$, say $p(\gamma_k|\mathbf{y})$, $k = 1, \ldots, m$. We will describe the method for the case that a scaled inverse chi-square prior distribution is assumed for the variance parameter σ^2. We let $\boldsymbol{\gamma}_{-k}$ denote the vector $\boldsymbol{\gamma}$ without the kth element, that is

$$\boldsymbol{\gamma}_{-k} = (\gamma_1, \ldots, \gamma_{k-1}, \gamma_{k+1}, \ldots, \gamma_m)^{\mathrm{T}}. \tag{5.23}$$

We also use the notation $\sum_{i\neq k}$ for the summation excluding the k indexed element of a sequence. Then using (5.18), the marginal posterior density $p(\gamma_k|\mathbf{y})$, for fixed γ_k, is given by

$$\begin{aligned} p(\gamma_k|\mathbf{y}) &= \int_{\mathbb{R}^{m-1}} p(\gamma_k, \boldsymbol{\gamma}_{-k}|\mathbf{y})\, d\boldsymbol{\gamma}_{-k} \\ &= \frac{\int_{\mathbb{R}^{m-1}} \exp\{\gamma_k(y_k+1) - e^{\gamma_k}\} \exp\{\sum_{i\neq k}^{m} \gamma_i(y_i+1) - \sum_{i\neq k}^{m} e^{\gamma_i}\}\, U(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}_{-k}}{\int_{\mathbb{R}^m} U(\boldsymbol{\gamma})\, I(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}} \\ &= \frac{\int_{\mathbb{R}^{m-1}} I(\gamma_k)\, U(\boldsymbol{\gamma})\, I(\boldsymbol{\gamma}_{-k})\, d\boldsymbol{\gamma}_{-k}}{\int_{\mathbb{R}^m} U(\boldsymbol{\gamma})\, I(\boldsymbol{\gamma})\, d\boldsymbol{\gamma}} \\ &= \frac{\mathrm{E}_I^*\{I(\gamma_k)\, U(\boldsymbol{\gamma})\}}{\mathrm{E}_I\{U(\boldsymbol{\gamma})\}}. \end{aligned} \tag{5.24}$$

Here, $U(\boldsymbol{\gamma})$ is given in (5.19) and $I(\gamma_k)$is given by

$$I(\gamma_k) = \exp\{\gamma_k(y_k+1) - e^{\gamma_k}\}. \tag{5.25}$$

The expectation E_I in the denominator is taken, as before, with respect to the joint distribution of m independent random variables $\gamma_1, \gamma_2, \ldots, \gamma_m$ where $e^{\gamma_i} \sim \mathrm{Ga}(y_i+1, 1)$. However, it is important to notice that in the numerator, γ_k is fixed and therefore, the expectation E_I^* corresponds to the distribution of the $(m-1)$ remaining random variables $\gamma_1, \ldots, \gamma_{k-1}, \gamma_{k+1}, \ldots, \gamma_m$, where again $e^{\gamma_i} \sim \mathrm{Ga}(y_i+1, 1)$.

Hence, we can estimate the marginal posterior density of γ_k using the importance sampling estimator

$$p^{\mathrm{IS}}(\gamma_k|\mathbf{y}) = \frac{\sum_{j=1}^{N} I(\gamma_{kj})\, U(\boldsymbol{\gamma}_j^*)}{\sum_{j=1}^{N} U(\boldsymbol{\gamma}_j)} \tag{5.26}$$

where

$$U(\boldsymbol{\gamma}_j) = e^{-\sum_{i=1}^{m}\gamma_{ij}} \left\{\nu\lambda + \sum_{i=1}^{m}(\gamma_{ij} - \bar{\gamma}_j)^2\right\}^{-(\frac{\nu+m-1}{2})},$$

and

$$\boldsymbol{\gamma}_j^* = (\gamma_{1j}, \ldots, \gamma_{k-1,j}, \gamma_k, \gamma_{k+1,j}, \ldots, \gamma_{mj})^{\mathrm{T}}$$

being independently simulated such that $e^{\gamma_{ij}} \sim \mathrm{Ga}(y_i+1, 1)$ for $i \neq k$, with γ_k kept fixed across the simulations $j = 1, \ldots, m$. Also, each $\boldsymbol{\gamma}_j$ is generated in the same manner, but without fixing γ_k.

We notice that under an appropriate Inv-$\chi^2(\nu, \lambda)$ hyperprior distribution for σ^2, the importance sampling approximation (5.26) can be computed with a finite standard error of simulation. Then, we can transform back to the marginal posterior density of θ_k, using

$$p(\theta_k|\mathbf{y}) = e^{-\gamma_k}\, p(\gamma_k|\mathbf{y}).$$

It is important to stress here that the importance sampling estimates described in this chapter may still be unreliable after a very large number of Monte Carlo simulations, even when the conditions for a finite standard error of simulation are met. In practice, the standard error of simulation may be considerably large for any practically feasible number of simulations as shown in the examples of Subsections 5.2.3 and 5.2.4 and also demonstrated in Geweke (1989). Furthermore, as emphasised earlier in the present chapter, the efficiency of the method often depends on the data sample size or relies on selecting an informative scaled inverse chi-square prior distribution for σ^2, with the appropriate degrees of freedom. Our empirical experience suggested that, often, in the absence of a sufficiently large sample, we need to increase the prior information included in the model. This results in less flexible modelling, and implies that the satisfactory performance of the method requires that, for the prior setting we should also consider the technical characteristics of the method, rather than clearly reflect our prior knowledge and beliefs.

The difficulties that we have encountered in implementing this particular importance sampling method for the full hierarchical analysis of the Poisson/log-normal model, may suggest that a different importance density should be chosen, depending on the specific data set and prior setting, or that an overall different approach should be adopted.

5.3 Markov chain Monte Carlo methods

In Section 5.2 we investigated the implementation of an importance sampling technique for simulating from the distribution of interest, that is the posterior distribution of the Poisson parameters $\theta_1, \theta_2, \ldots, \theta_m$. Importance sampling is a noniterative simulation method, in the sense that it only employs a sufficiently large sequence of generated values from a single distribution. The method requires that this distribution is an adequately good approximation to the target posterior distribution, otherwise it leads to poor estimation.

When no such approximating distribution is available for simulation, we can sample from an appropriate Markov chain that converges to the desired posterior distribution. The idea is to create a Markov process whose stationary distribution is the distribution of interest, and run this stochastic process for long enough, so that the distribution of the current draws is adequately close to the stationary distribution. We can then sample the converged process to obtain the inferences of interest. This is an iterative procedure, since it consists of drawing values from an iterated sequence of distributions which eventually converges to the desired

target distribution. The method, which has become known under the general term Markov chain Monte Carlo (MCMC), originates in the context of statistical physics in Metropolis *et al.* (1953) and was generalised by Hastings (1970). Geman and Geman (1984) introduced a special case of the more general methods presented by the previous authors, namely the Gibbs sampler, again in the context of statistical physics. Gelfand and Smith (1990) and Gelfand *et al.* (1990) showed the potential of MCMC methodology, and mainly that of the Gibbs sampler, in conventional statistical problems. Smith and Roberts (1993) and Tierney (1994) discuss its implementation to Bayesian computation, while also providing the necessary theoretical framework. Recent books from Gelman *et al.* (1995), Gilks, Richardson, and Spiegelhalter (1996) and Carlin and Louis (1996) provide a thorough presentation of the MCMC methodology.

5.3.1 Markov chain theory

Before we further discuss the construction of an appropriate Markov chain for simulation from the distribution under consideration, we provide a brief account of the main aspects of the theory of Markov chains, that will be used later in this section.

Suppose we have a discrete time stochastic process $\{X_0, X_1, X_2, \dots\}$, with the property that given the history of the process, the future development may depend on the present, but does not depend on the past. Formally, we can write

$$\mathrm{P}(X_{n+1} \in A | X_n = s, X_{n-1} \in A_{n-1}, X_{n-2} \in A_{n-2}, \dots, X_0 \in A_0) = \mathrm{P}(X_{n+1} \in A | X_n = s), \quad (5.27)$$

for any sets $A_0, A_1, A_2, \dots, A_{n-1}, A \subset S$ and $s \in S$, where S is the state space of the process. A stochastic process with property (5.27) is known as a Markov chain. Here, for presentation purposes, we will assume that S is a discrete state space. Analogous results hold for general state spaces. We also assume that the probability $\mathrm{P}(X_{n+1} = y | X_n = x)$ for any $x, y \in S$ does not depend on n, in which case the chain is said to be homogeneous. Then, we can define $\mathrm{P}(x, y)$, the one step transition kernel, or probability, of the chain as

$$\mathrm{P}(x, y) = \mathrm{P}(X_{n+1} = y | X_n = x), \qquad \text{for all } x, y \in S. \quad (5.28)$$

For all $x \in S$, $\mathrm{P}(x, \cdot) = \mathrm{P}(\cdot | X_n)$ defines a probability distribution over S, and thus, $\mathrm{P}(x, y) \geq 0$ for all $x, y \in S$ and $\sum_y \mathrm{P}(x, y) = 1, \forall x \in S$.

The n-step transition probability, that is the probability that the chain is at state y in exactly n steps given that the system starts in x, is denoted by

$$\mathrm{P}^n(x, y) = \mathrm{P}(X_n = y | X_0 = x), \quad (5.29)$$

and subject to the regularity conditions stated below, $\mathrm{P}^n(x, y)$ will converge to a unique stationary distribution $\pi(\cdot)$, also known as the invariant distribution, which does not depend on n or x. This means that the Markov chain will eventually 'forget' its starting position.

The conditions for the distribution of X_n to converge to the unique stationary distribution $\pi(\cdot)$ are that the Markov chain $\{X_0, X_1, X_2, \dots\}$ is irreducible, aperiodic and positive recurrent (Roberts, 1995). Irreducibility implies that all states of the Markov chain communicate with each other; aperiodicity ensures that the chain does not oscillate between some set of states in regular times; and positive recurrence ensures that there exists a stationary distribution $\pi(\cdot)$, such that if the initial probabilities of the chain being in state $0, 1, 2, \dots$ are given by $\pi(\cdot)$, then these probabilities remain unaltered at all subsequent times. Formally we have the following definitions. First we use T_{xx} to denote the time of the first return to state x, i.e. $\mathrm{T}_{xx} = \min\{n \geq 1 : X_n = x | X_0 = x\}$.

Definition 5.1. *The chain X_n is called irreducible if for all states $x, y \in S$, there exists $n \geq 1$ such that $P^n(x, y) > 0$.*

Definition 5.2. *The chain X_n is called aperiodic if for all states $x \in S$, the largest integer t, such that all the times at which the chain returns to x are multiples of t, is equal to 1.*

Definition 5.3. *An irreducible chain X_n is said to be recurrent if for all states $x \in S$, $P(T_{xx}) < \infty = 1$.*

Definition 5.4. *An irreducible recurrent Markov chain X_n is called positive recurrent if for all states $x \in S$, $E(T_{xx}) < \infty$. Equivalently, as stated in Tierney (1995), X_n is positive recurrent iff there exists a stationary probability distribution $\pi(\cdot)$ for X_n, that is iff there exists $\pi(\cdot)$ such that*

$$\sum_x \pi(x) P(x, y) = \pi(y), \qquad \text{for all } y \in S. \tag{5.30}$$

Definition 5.5. *A Markov chain $\{X_0, X_1, X_2, \dots\}$ is called time reversible if it is positive recurrent with invariant distribution $\pi(\cdot)$ and*

$$\pi(x) P(x, y) = \pi(y) P(y, x), \qquad \text{for all } x, y \in S.$$

From Definition 5.4 it follows that if X_n is a positive recurrent Markov chain, its stationary distribution is the unique probability distribution $\pi(\cdot)$ satisfying (5.30). Furthermore, if X_n is also aperiodic, we say that X_n is ergodic, and the following theorem holds (Tierney, 1994).

Theorem 5.1. *For an irreducible ergodic Markov chain X_n with transition probabilities $P(x, y)$,*

(i) $P^n(x, y) \to \pi(y)$ *as* $n \to \infty$, $\forall x, y \in S$

(ii) let $h(x)$ be a real valued function and $\hat{h}_N = \frac{1}{N}\sum_{n=1}^{N} h(X_n)$ be the so-called ergodic average of $h(x)$. Then, if $E_\pi\{h(x)\} < \infty$, where the expectation is taken with respect to the stationary distribution $\pi(\cdot)$ of the chain,

$$\hat{h}_N \to E_\pi\{h(x)\} \text{ as } n \to \infty \text{ with probability } 1.$$

The second result of Theorem 5.1 is equivalent to the law of large numbers and allows us to estimate consistently various characteristics of the invariant distribution $\pi(\cdot)$, using the dependent realisations of the Markov chain.

5.3.2 Markov chain Monte Carlo

We now discuss how we can implement the theoretical properties of Markov chains outlined in the preceding section, to simulate from a distribution of interest. Suppose that we wish to generate a random variable X, having a probability distribution $\pi(x)$. As already mentioned, if we can neither sample $\pi(x)$ directly, nor we can use a noniterative simulation scheme, we may attempt to create a Markov chain $\{X_0, X_1, X_2, \dots\}$ having $\pi(\cdot)$ as its stationary distribution. We could then sample the later stages of the chain, after we have allowed it to run long enough to ensure that it has converged to its invariant distribution.

The construction of such a Markov chain first requires that the conditions of irreducibility and aperiodicity of the chain are met, as formally shown by Smith and Roberts (1993) and Tierney (1994). The condition of positive recurrence in Theorem 5.1 follows from the fact that we already know that $\pi(\cdot)$, the stationary distribution of the chain under cosideration exists, that being the distribution of the random variable X. Therefore, from Definition 5.4, the chain is positive recurrent, with its unique invariant distribution $\pi(\cdot)$ satisfying (5.30). Then, we only need to define the chain transition probabilities $\mathrm{P}(x, y)$ involved in (5.30), and given that the chain is irreducible and aperiodic, from Theorem 5.1 we deduce that $\mathrm{P}^n(x, y) \to \pi(y)$ as $n \to \infty$. That is, we only need to find the transition kernel $\mathrm{P}(x, y)$ of the Markov chain and allow a sufficient number of transition steps to ensure that $\mathrm{P}^n(x, y)$ has converged to $\pi(\cdot)$. The problem of defining a transition kernel $\mathrm{P}(x, y)$ to satisfy (5.30) is facilitated by using reversible chains. If we can obtain a probability $\mathrm{P}(x, y)$ such that the chain has the property of reversibility, i.e.

$$\pi(x)\mathrm{P}(x, y) = \pi(y)\mathrm{P}(y, x), \qquad \text{for all } x, y \in S, \tag{5.31}$$

then $\mathrm{P}(x, y)$ gives the desired transition kernel. This follows from the fact that summing both sides of (5.31) over x will give (5.30). Hence, the remaining question is how to construct a transition kernel $\mathrm{P}(x, y)$ that satisfies the reversibility condition (5.31). The following section deals with this problem.

5.3.3 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm, as given by Hastings (1970), is a generalisation of the technique proposed by Metropolis *et al.* (1953). Chib and Greenberg (1995) provide a simple and intuitive description of the algorithm and a detailed presentation of the method and the underlying theory can be found in Tierney (1994).

We now return to the problem of the construction of the transition probability $\mathrm{P}(x, y)$ in such a way that the reversibility condition (5.31) holds. For the Metropolis-Hastings algorithm we update the irreducible and aperiodic Markov chain described in the preceding section as follows: when $X_n = x$ we generate a candidate variate y from a proposal distribution $q(x, y)$ and accept the new value with probability $\alpha(x, y)$, which will be defined later. Then, the probability that the Markov chain moves from state x to state y, when $x \neq y$, is given by

$$\mathrm{P}(x, y) = q(x, y)\ \alpha(x, y), \qquad \text{if } x \neq y.$$

However, there is also a non-zero probability that the chain remains in state x, which can be expressed as the probability of rejection of all other possible candidates y, that is

$$\mathrm{P}(x, x) = 1 - \sum_{y \neq x} q(x, y)\alpha(x, y).$$

By merging the two above cases, the transition probabilities of the Markov chain are given by

$$\mathrm{P}(x, y) = q(x, y)\ \alpha(x, y) + I(y = x)\ \{1 - \sum_{z \neq x} q(x, z)\alpha(x, z)\}, \qquad \forall x, y \in S, \quad (5.32)$$

where $I(\cdot)$ denotes the indicator function, taking the value 1 when $y = x$ and the value 0 otherwise. We want the reversibility property (5.31) to be valid when the transition probabilities are given by (5.32). Clearly, property (5.31) is satisfied when $x = y$. If $x \neq y$, using the first term of (5.32) we notice that the reversibility condition can be written as

$$\pi(x)\ q(x, y)\ \alpha(x, y) = \pi(y)\ q(y, x)\ \alpha(y, x), \qquad \forall x, y \in S \text{ with } x \neq y. \quad (5.33)$$

Now, if we define the acceptance probabilities $\alpha(x, y)$ to be

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)\ q(y, x)}{\pi(x)\ q(x, y)}\right\}, \tag{5.34}$$

then (5.33) is satisfied, since if $\alpha(x, y) = \frac{\pi(y)\ q(y,x)}{\pi(x)\ q(x,y)}$, then $\alpha(y, x) = 1$ and thus (5.33) holds. Also, if $\alpha(x, y) = 1$, then $\alpha(y, x) = \frac{\pi(x)\ q(x,y)}{\pi(y)\ q(y,x)}$ and again (5.33) follows.

The described method is known as the Metropolis-Hastings algorithm, and we have demonstrated that it produces a Markov chain which has the target distribution $\pi(x)$ as its stationary distribution. The transition kernel of the chain is given in (5.32), where $q(x, y)$ is a proposal distribution and $\alpha(x, y)$ in (5.34) is the acceptance probability for the candidate points which are simulated from $q(x, y)$. Although we have presented the method using Markov chains with a discrete state space, the same results hold for general state chains, meaning that X may also be a continuous random variable. The proposal distribution $q(x, y)$ may depend on the current point X_n and, in theory, it can be any probability distribution from which we can simulate. However, different forms of $q(\cdot, \cdot)$ may lead to faster convergence of the Markov chain to the desired target distribution.

We can now give an outline of the Metropolis-Hastings algorithm. Suppose that we want to simulate a random variable X from a distribution of interest denoted by $\pi(\cdot)$. We proceed as follows:

ALGORITHM 5.1: *THE METROPOLIS-HASTINGS METHOD.*

1. *Initialise with a starting point* $X^{(0)}$
2. *For* $t = 1, 2, 3, \ldots$ *(until convergence):*
   (a) *Sample a candidate point* Y *from a proposal distribution at time* t, $q(\cdot|X^{(t-1)})$
   (b) *Calculate the ratio* $r = \frac{\pi(Y)\ q(X^{(t-1)}|Y)}{\pi(X^{(t-1)})\ q(Y|X^{(t-1)})}$
   (c) *Set* $X^{(t)} = \begin{cases} Y & \text{with probability } min(1, r) \\ X^{(t-1)} & \text{otherwise} \end{cases}$

The method requires the calculation of the ratios r and the ability to draw random variables from the proposal distribution $q(\cdot, \cdot)$. Therefore, for efficiency reasons, the proposal distribution should be chosen to be easy to evaluate, yet to lead to rapid mixing, that is it to move rapidly around the support of $\pi(\cdot)$.

In practice, it is often difficult to satisfy both of these requirements simultaneously, and therefore a trade-off between efficient evaluation and rapid mixing is preferred.

We notice that when the proposal distribution is symmetric, that is when $q(x,y) = q(y,x)$ for all x, y, e.g. when a normal distribution centred at the 'old' point is chosen, the acceptance probability simplifies to $\alpha(x,y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$ leading to the Metropolis algorithm, which is the method originally proposed by Metropolis et *et al.* (1953).

5.3.4 The Gibbs sampler

The Gibbs sampler is a special case of the Metropolis-Hastings algorithm. It was first presented by Geman and Geman (1984), and since the work of Gelfand and Smith (1990), it has been widely used in a vast range of statistical problems. Together with the entire MCMC methodology, the Gibbs sampler in particular has greatly facilitated Bayesian computation. The papers from Gelfand *et al.* (1990) and Zeger and Karim (1991) have contributed towards this direction. Casella and George (1992) present a clear explanation of the method.

Consider now that we want to simulate a vector $\mathbf{X} = (X_1, \ldots, X_m)^{\mathrm{T}}$ from an m-dimensional distribution of interest, denoted by $\pi(\cdot)$. The Metropolis-Hastings algorithm, can still be applied as presented in Subsection 5.3.3, by updating the whole vector $\mathbf{X}$ in a single block. However, we may also update the m components of $\mathbf{X}$ separately, one at a time, or in groups. Suppose that we update a single component X_i of $\mathbf{X}$ at a time, following the natural order $i = 1, 2 \ldots, m$. Then, each iteration of the Metropolis-Hastings algorithm comprises m updating steps. We let $\mathbf{X}_{-i}$ denote the vector $\mathbf{X}$ without its ith component, that is

$$\mathbf{X}_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m).$$

At step i of iteration t we generate a candidate variate Y from a proposal distribution

$$q_i\left(Y_i | X_1^{(t)}, \ldots, X_{i-1}^{(t)}, X_i^{(t-1)}, X_{i+1}^{(t-1)}, \ldots, X_m^{(t-1)}\right) = q_i\left(Y_i | X_i^{(t-1)}, \mathbf{X}_{-i}^{(t-1)}\right),$$

where $\mathbf{X}_{-i}^{(t-1)}$ represents the vector of all components of $\mathbf{X}$, except X_i, at their current values, i.e.

$$\mathbf{X}_{-i}^{(t-1)} = \left(X_1^{(t)}, \ldots, X_{i-1}^{(t)}, X_{i+1}^{(t-1)}, \ldots, X_m^{(t-1)}\right),$$

following from the fact that $(i-1)$ components have already been updated at step i. Then, the acceptance probability (5.34) for the ith component of $\mathbf{X}$ at

iteration t becomes

$$\alpha(X_i, Y_i) = \min\left\{1, \frac{\pi\left(Y_i|\mathbf{X}_{-i}^{(t-1)}\right)\ q_i\left(X_i^{(t-1)}|Y_i, \mathbf{X}_{-i}^{(t-1)}\right)}{\pi\left(X_i^{(t-1)}|\mathbf{X}_{-i}^{(t-1)}\right)\ q_i\left(Y_i|X_i^{(t-1)}, \mathbf{X}_{-i}^{(t-1)}\right)}\right\}, \tag{5.35}$$

leading to the so-called single-component Metropolis-Hastings algorithm. If Y_i is accepted, then we set $X_i^{(t)} = Y_i$, otherwise $X_i^{(t)} = X_i^{(t-1)}$. Notice that only the ith component of $\mathbf{X}$ is changed at step i of each iteration of the algorithm.

The distribution $\pi(X_i|\mathbf{X}_{-i})$ is known as the full conditional distribution of X_i given all the remaining components of $\mathbf{X}$. The Gibbs sampler is a single-component Metropolis-Hastings algorithm which updates the ith element of $\mathbf{X}$ at time t according to the full conditional distribution of X_i, that is it employs a proposal distribution of the form

$$q_i\left(Y_i|X_i^{(t-1)}, \mathbf{X}_{-i}^{(t-1)}\right) = \pi\left(Y_i|\mathbf{X}_{-i}^{(t-1)}\right). \tag{5.36}$$

Therefore, substitution of (5.36) in (5.35), implies that the Gibbs sampler always accepts a candidate point generated from the full conditional distribution of the component of $\mathbf{X}$ that is currently being updated. This means that at iteration t, each element X_i is updated with a value generated from its conditional distribution , given the latest value of the other components, which is the iteration t value for the already changed elements of $\mathbf{X}_{-i}$ and the iteration $(t-1)$ value for the remaining elements. The algorithm can be outlined as follows:

ALGORITHM 5.2: *THE GIBBS SAMPLING METHOD.*

1. *Choose starting values* $\mathbf{X}^{(0)} = \left(X_1^{(0)}, \ldots, X_m^{(0)}\right)$
2. *For* $t = 1, 2, 3, \ldots$ *(until convergence):*
   - (a) *At iteration t, take as input the point* $\mathbf{X}^{(t-1)}$
   - (b) *Generate* $X_1^{(t)}$ *from* $\pi\left(X_1|X_2^{(t-1)}, \ldots, X_m^{(t-1)}\right)$

     $\vdots$

     Generate $X_i^{(t)}$ *from* $\pi\left(X_i|X_1^{(t)}, \ldots, X_{i-1}^{(t)}, X_{i+1}^{(t-1)}, \ldots, X_m^{(t-1)}\right)$

     $\vdots$

     Generate $X_m^{(t)}$ *from* $\pi\left(X_1|X_1^{(t)}, \ldots, X_{m-1}^{(t)}\right)$
   - (c) *Set* $\mathbf{X}^{(t)} = \left(X_1^{(t)}, \ldots, X_m^{(t)}\right)$

The updating order for the m components of $\mathbf{X}$ does not need to be fixed. More-

over, not all the components have to be altered at each iteration. This strategy would lead to a blocking scheme.

Clearly, the method requires that the full conditional distributions of all the elements of $\mathbf{X}$ can be fully specified and are available for sampling. Under mild conditions (Besag, 1974), the specification of all full conditional distributions uniquely determines the full joint distribution, and hence all marginal distributions. Then, as demonstrated for the more general Metropolis-Hastings method, under the conditions of irreducibility and aperiodicity, as $t \to \infty$, $\left(X_1^{(t)}, \ldots, X_m^{(t)}\right)$ will converge in distribution to $\mathbf{X} \sim \pi(\mathbf{X})$. Hence, for $i = 1, \ldots, m$, X_i converges in distribution to $X_i \sim \pi(X_i)$ as $t \to \infty$, and therefore a sample from all the marginal distributions $\pi(X_i)$ is available without any further effort. This can be used to obtain estimates of various characteristics of the distribution of X_i. We notice here that the dependence of the sample will not usually be a problem, since applicability of Theorem 5.1 implies that we can still use ergodic averages to estimate expectations of real valued functions with respect to $\pi(\cdot)$.

In the case that the estimation of the marginal probability density function $\pi(X_i)$ is of interest, we may exploit the known form of the full conditional distribution $\pi(X_i|\mathbf{X}_{-i})$ to obtain an efficient estimator. Since the probability density function of X_i can be expressed as $\pi(X_i) = \int \pi(X_i|\mathbf{X}_{-i})\pi(\mathbf{X}_{-i})\, d\mathbf{X}_{-i}$, we can estimate it using

$$\hat{\pi}(X_i) = \frac{1}{N}\sum_{j=1}^{N} \pi\left(X_i|\mathbf{X}_{-i}^{(j)}\right), \tag{5.37}$$

where N is the size of the simulated sample. Gelfand and Smith (1990) suggest this estimator and they refer to it as the 'Rao-Blackwellised' estimator, since based on the Rao-Blackwell theorem, it offers some variance reduction when compared to the more usual kernel estimation method. Similarly, if the inference concerns the mean of the marginal distribution $\pi(X_i)$, we may use the estimator

$$\hat{\mathrm{E}}(X_i) = \frac{1}{N}\sum_{j=1}^{N} \mathrm{E}\left(X_i|\mathbf{X}_{-i}^{(j)}\right). \tag{5.38}$$

It is important to stress that both estimators (5.37) and (5.38) require knowledge of the normalised form of the full conditional distribution of the parameter of interest. When this is not available, a kernel estimator and an ergodic average can be employed respectively.

5.3.5 Implementation issues

The implementation of the Gibbs sampler raises various practical issues regarding the choice of the starting values of the algorithm, the simulation from the full

conditional distributions, the convergence of the Markov chain, the dependence of the generated sample, and the use of the output.

As long as the chain is irreducible, the selection of the initial values is not important, provided that we can run the chain long enough so that it converges to the target distribution. However, when the aim is to obtain adequately good estimates after a relatively small number of iterations, as would the case be if a frequency properties study was to be conducted, it would be useful to start the chain at initial points which are not too far from good guesses of the parameters under consideration. On the other hand, Gelman and Rubin (1992) favour the use of multiple chains and therefore suggest that the starting points should be widely scattered through the support of the target distribution, so that all main parts of the distribution are represented in the simulations. In either case a crude method for obtaining some starting values of the distribution of interest should be available.

The theoretical convergence of the constructed chain to the target distribution $\pi(\cdot)$ does not guarantee that we will obtain a variate from $\pi(\cdot)$ after a finite number of iterations. Detecting convergence to the stationary distribution of the Markov chain is one of the main problems with MCMC implementation and is still an active area of research. Some theoretical results are available (e.g. Smith and Roberts, 1993), but they are usually difficult to apply. In practice, the assessment of the convergence of an MCMC algorithm relies on the statistical analysis of the output from the chain. Several convergence diagnostics have been proposed, including the following: Geweke (1992) suggests a method based on time series analysis to test convergence of ergodic averages; Gelman and Rubin (1992) use multiple chains to compare the variances between and within the chains; Zellner and Min (1995), for the case of the Gibbs sampler, propose a statistic which examines the agreement between two versions of the joint parameter distribution, which are based on the conditional distributions of a suitable partition of the parameter vector; Raftery and Lewis (1992) suggest the use of an approximately independent sample from a single long chain to estimate the variance of a computed characteristic. It is also widely suggested to check the autocorrelations in the output to detect a possible strong dependence between successive simulations.

There are advantages and disadvantages with all the convergence diagnostics that have appeared in the literature. Cowles and Carlin (1996) provide a thorough comparison of several methods in their review paper. In most applications, the convergence of an MCMC algorithm may also be informally assessed by visual inspection of the chain output. In practice, the problem is how to determine the so called burn-in period, that is a number of, say k iterations, after which

the algorithm produces a variate that has approximately the distribution of interest. Gelfand and Smith (1990) monitor the density estimates of components of the parameter vector, produced by k-batches of independent implementations of the method. They compare these estimates for increasing values of k, and they suggest stopping at that k value, for which the densities appear to be indistinguishable. The same idea can be employed by plotting ergodic averages of functions of parameters of interest and visually checking convergence. As Gelman and Rubin (1992) suggest, it is again useful to start a number of chains at overdispersed points and monitor their output simultaneously, in order to check whether and when all chains produce the same output. It is also important to assess the convergence of several, if not all, the distributions of the model parameters.

Once the length k of the burn-in period has been specified, we can discard the first k simulated values to diminish the effect of the starting position of the chain, and sample from the target distribution. Then, a question that often arises is how to deal with the dependence of the sampled values. As mentioned before, this dependence will not affect most inferences, due to the exploitation of ergodic results. Therefore, if a sample of size N is required, we can use the last N simulations of a single long chain. This approach was strongly backed by Geyer (1992). Nevertheless, Gelman and Rubin (1992) propose the use of the last $\frac{N}{l}$ values produced by each of l independent chains. However, if an independent N-sample is desired, one can either take the $k+1$ iteration of N independent chains, as suggested by Gelfand and Smith (1990), or utilise a long chain to sample every rth simulation draw after the burn-in period, for N times. If r is large enough, an approximately independent sample will be obtained.

Finally, as far the Gibbs sampler is concerned, an additional implementation issue occurs when one or more of the full conditional distributions cannot be specified in closed form, and therefore simulation from them is not possible. If one does not choose an overall different MCMC algorithm, various approaches have been proposed to overcome this problem. One idea is to use rejection sampling to simulate from the conditional distributions of nonclosed form. Zeger and Karim (1991) and Carlin and Gelfand (1991) employ rejection sampling techniques with normal, t or split-t envelope functions. Wakefield *et al.* (1991) suggested a ratio-of-uniforms version of the rejection algorithm, and Gilks and Wild (1992) developed a popular adaptive rejection algorithm, forming an envelope function by intersecting the tangent or secant lines at pre-chosen points of the density of interest, as long as the latter is log-concave. An alternative approach is considered by Ritter and Tanner (1992), who recommended a generalised inversion method (e.g., Devroye, 1986), utilising a discrete approximation to the full conditional cumulative

density function, obtained from a grid-based evaluation of the full conditional probability density function. Another alternative of increasing popularity consists of combining the Gibbs sampler with other MCMC techniques, an approach that leads to the so-called hybrid MCMC algorithms (Tierney 1994).

5.4 Gibbs sampling for the Poisson/log-normal model

We will now investigate how we can implement the MCMC methodology to the full Bayesian hierarchical analysis of the Poisson/log-normal model (4.1a), or (4.1b). We will first consider the Gibbs sampling algorithm. Zeger and Karim (1991) discuss the application of the Gibbs sampling approach to the analysis of generalised linear models, while George, Makov and Smith (1994) and Gelfand and Smith (1990) illustrate the method in the case of the conjugate Poisson/gamma formulation. Carlin and Gelfand (1991) also include an analysis of a Poisson/log-t model, in addition to the conjugate case. All these authors employ the Gibbs sampler combined with various rejection sampling schemes, wherever this is necessary. Damien, Wakefield and Walker (1999) adopt an auxiliary variable technique to analyse a Poisson/log-normal model using Gibbs sampling. Tierney (1994) considers the implementation of Metropolis-Hastings and Gibbs sampling methods, including hybrid MCMC techniques, in the case of Poisson/gamma and Poisson/log-normal models.

5.4.1 Derivation of the full conditional distributions

As stressed in Subsection 5.3.4, for the implementation of the Gibbs sampler we first need to determine the full conditional distributions of all the unknown model parameters. In the Bayesian context, the distribution under consideration is the posterior distribution of the parameters of interest and thus, for the Poisson/log-normal models (4.1a) and (4.1b), the Gibbs sampler requires that we provide the full conditional posterior distributions of the Poisson means $\theta_1, \theta_2, \ldots, \theta_m$, or the equivalent for $\gamma_1, \gamma_2, \ldots, \gamma_m$, and those of the hyperparameters μ and σ^2, given all the other parameters.

Full conditional distributions under model (4.1a)

We first derive the full conditional posterior distributions of the parameters for model (4.1a), under the parametrisation $\gamma_i = \log(\theta_i)$, for $i = 1, \ldots, m$. In Subsection 5.2.1 we gave the joint posterior density of the parameters $\boldsymbol{\gamma}, \mu$ and σ^2

as

$$p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}m} \exp\left[\sum_{i=1}^{m}\left\{\gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2\right\}\right] \tag{5.39}$$

$$\propto (\sigma^2)^{-\frac{1}{2}m} \exp\left\{\sum_{i=1}^{m}\gamma_i y_i - \sum_{i=1}^{m} e^{\gamma_i} - \frac{1}{2}\sigma^{-2}\sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}$$

$$\exp\left\{-\frac{1}{2}\sigma^{-2}m\ (\mu - \bar{\gamma})^2\right\}. \tag{5.40}$$

Then, from (5.40) follows that

$$p(\mu|\boldsymbol{\gamma}, \sigma^2, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\sigma^{-2}m\ (\mu - \bar{\gamma})^2\right\},$$

implying that, given $\boldsymbol{\gamma}$ and σ^2, the full conditional posterior distribution of μ is normal with mean $\bar{\gamma}$ and variance $\sigma^2\ m^{-1}$, i.e.

$$\mu \mid \boldsymbol{\gamma}, \sigma^2, \mathbf{y} \sim N\left(\bar{\gamma}, \frac{\sigma^2}{m}\right). \tag{5.41}$$

Also, from (5.39) we can see that the full conditional posterior distribution of σ^2 can be written as

$$p(\sigma^2|\boldsymbol{\gamma}, \mu, \mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}m} \exp\left\{-\frac{1}{2}\sigma^{-2}\ \sum_{i=1}^{m}(\gamma_i - \mu)^2\right\}, \tag{5.42}$$

and therefore, the full conditional posterior distribution of $\frac{\sum_{i=1}^{m}(\gamma_i-\mu)^2}{\sigma^2}$ is given by

$$p\left(\frac{\sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2} \,\middle|\, \boldsymbol{\gamma}, \mu, \mathbf{y}\right) \propto$$

$$\left\{\frac{\sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2}\right\}^{\frac{m-2}{2}-1} \exp\left\{-\frac{1}{2}\ \frac{\sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2}\right\}, \tag{5.43}$$

and hence, given $\boldsymbol{\gamma}$ and μ the the full conditional posterior distribution of $\frac{\sum_{i=1}^{m}(\gamma_i-\mu)^2}{\sigma^2}$ is chi-square with $m-2$ degrees of freedom. We write

$$\frac{\sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2} \,\Big|\, \boldsymbol{\gamma}, \mu, \mathbf{y} \sim \chi^2_{m-2}. \tag{5.44}$$

Clearly, (5.44) can be used to obtain variates from the full conditional posterior distribution of σ^2, as will be described in a subsequent section.

As far as the full conditional posterior distribution of the parameter vector $\boldsymbol{\gamma}$ is concerned, (5.39) implies that given the hyperparameters μ and σ^2, the parameters $\gamma_1, \gamma_2, \ldots, \gamma_m$, are independent, and their full conditional posterior distribution is given by

$$p(\gamma_i|\mu, \sigma^2, \mathbf{y}) \propto \exp\left\{\gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\ \sigma^{-2}(\gamma_i - \mu)^2\right\}, \qquad i = 1 \ldots, m. \tag{5.45}$$

Full conditional distributions under model (4.1b)

In the case that a Inv-$\chi^2(\nu, \lambda)$ prior distribution is assumed for the variance parameter σ^2, independently from a uniformly distributed μ, model (4.1b) implies that the joint posterior density of $(\boldsymbol{\gamma}, \mu, \sigma^2)$ is now given by

$$p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-(\frac{\nu+m}{2}+1)} \exp\left\{\sum_{i=1}^{m} \gamma_i y_i - \sum_{i=1}^{m} e^{\gamma_i} - \frac{1}{2}\sigma^{-2}\sum_{i=1}^{m}(\gamma_i - \bar{\gamma})^2\right\}$$
$$\exp\left\{-\frac{1}{2}\sigma^{-2}m\,(\mu - \bar{\gamma})^2\right\} \exp\left\{-\frac{\nu\lambda}{2\sigma^2}\right\}. \tag{5.46}$$

It is easy to verify that the full conditional distributions of μ and γ_i, $i = 1, \ldots, m$, are the same as for model (4.1a), and therefore are given by (5.41) and (5.45) respectively. However, as far as σ^2 is concerned, the full conditional distribution in (5.42) now becomes

$$p(\sigma^2|\boldsymbol{\gamma}, \mu, \mathbf{y}) \propto (\sigma^2)^{-(\frac{\nu+m}{2}+1)} \exp\left\{-\frac{1}{2}\,\frac{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2}\right\},$$

and transforming for $\frac{\nu\lambda+\sum_{i=1}^{m}(\gamma_i-\mu)^2}{\sigma^2}$ we obtain

$$p\left(\frac{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2} \,\middle|\, \boldsymbol{\gamma}, \mu, \mathbf{y}\right) \propto$$
$$\left\{\frac{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2}\right\}^{\frac{\nu+m}{2}-1} \exp\left\{-\frac{1}{2}\,\frac{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2}\right\}.$$

It is therefore obvious that the full conditional posterior distribution of $\frac{\nu\lambda+\sum_{i=1}^{m}(\gamma_i-\mu)^2}{\sigma^2}$ given above, is chi-square with $\nu + m$ degrees of freedom. That is,

$$\frac{\nu\lambda + \sum_{i=1}^{m}(\gamma_i - \mu)^2}{\sigma^2} \,\Big|\, \boldsymbol{\gamma}, \mu, \mathbf{y} \sim \chi^2_{\nu+m}. \tag{5.47}$$

5.4.2 Sampling from the full conditional distributions

The full conditional posterior distributions of μ and σ^2 are derived in standard form, and therefore simulation from them is straightforward. For both model specifications, the full conditional posterior distribution of the mean hyperparameter μ is the normal distribution given in (5.41), and sampling from it requires little effort. To generate σ^2 variates according to the chi-square distribution (5.44) or (5.47), we can draw a random variable, say Y, from a χ^2_{m-2} or $\chi^2_{\nu+m}$ distribution, and set $\sigma^2 = \frac{\sum_{i=1}^{m}(\gamma_i-\mu)^2}{Y}$ or $\sigma^2 = \frac{\nu\lambda+\sum_{i=1}^{m}(\gamma_i-\mu)^2}{Y}$ respectively. In the applications throughout this thesis, in order to simulate a χ^2 variate, we utilise the general result that the χ^2_k distribution is a special case of a $\text{Ga}(\frac{k}{2}, \frac{1}{2})$ distribution.

However, as shown by (5.45), the full conditional posterior distribution of the Poisson means does not take a standard form, and therefore we cannot sample from it directly. Several suggestions of how to overcome this problem appear in the papers cited in the beginning of Section 5.4. We will present an alternative approach, which leads to approximate simulations from (5.45). Also, based on this approximate method, we will present a hybrid MCMC technique to obtain the desired simulations.

5.4.3 A mixture approximation to $p(\theta_i|\mu, \sigma^2, \mathbf{y})$

The full conditional posterior distribution of the components of the parameter vector $\boldsymbol{\gamma}$ is given in (5.45). Equivalently, from the definition of the model (4.1a), the full conditional posterior distribution of $\theta_i = e^{\gamma_i}$ given μ and σ^2, takes the form

$$p(\theta_i|\mu, \sigma^2, \mathbf{y}) \propto \theta_i^{y_i}\, e^{-\theta_i}\, \exp\left\{-\frac{1}{2}\,\sigma^{-2}\,(\log\theta_i - \mu)^2 - \log\theta_i\right\}, \tag{5.48}$$

for $i = 1, \ldots, m$.

The form of the probability density function (5.48) suggests that we may attempt to approximate the full conditional posterior distribution of θ_i with a mixture of a gamma and a log-normal distribution. The idea is to obtain an accurate approximation, by matching the three first moments of the original and the approximate distribution. In order to be able to do that, we first need to compute the moments of the nonclosed full conditional posterior distribution of θ_i in (5.48). The difficulty in doing so, occurs from the fact that the chosen method will have to be incorporated in the iterative scheme of the MCMC algorithm, and hence any computationally intensive technique should be avoided. We will discuss suitable methods in the next section.

Once the moments of the original distribution have been obtained, we choose the parameters of a $\mathrm{Ga}(a, b)$ and a $\mathrm{LN}(\delta, \tau^2)$ distribution, in a way such 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 posterior distribution. This is easily achieved by separately solving the following two sets of equations:

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

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

for a, b, and

$$\begin{aligned}\mathrm{E}(\theta_i|\mu,\sigma^2,\mathbf{y}) &= e^{\delta+\frac{1}{2}\tau^2}\\ \mathrm{var}(\theta_i|\mu,\sigma^2,\mathbf{y}) &= \left(e^{\delta+\frac{1}{2}\tau^2}\right)^2\ \left(e^{\tau^2}-1\right).\end{aligned}$$

for δ and τ^2. This will give

$$\begin{aligned}a &= \frac{\mathrm{E}^2(\theta_i|\mu,\sigma^2,\mathbf{y})}{\mathrm{var}(\theta_i|\mu,\sigma^2,\mathbf{y})}\\ b &= \frac{\mathrm{E}(\theta_i|\mu,\sigma^2,\mathbf{y})}{\mathrm{var}(\theta_i|\mu,\sigma^2,\mathbf{y})}\end{aligned} \tag{5.49}$$

for the parameters of the gamma distribution, and

$$\begin{aligned}\delta &= \log\{\mathrm{E}(\theta_i|\mu,\sigma^2,\mathbf{y})\} - \frac{\tau^2}{2}\\ \tau^2 &= \log\left\{1+\frac{\mathrm{var}(\theta_i|\mu,\sigma^2,\mathbf{y})}{\mathrm{E}^2(\theta_i|\mu,\sigma^2,\mathbf{y})}\right\}\end{aligned} \tag{5.50}$$

for the log-normal part of the mixture distribution. Using the parameters in (5.49) and (5.50) we can also derive the higher order moments for the log-normal and gamma distribution. The third moment of the log-normal distribution is given by $\exp(3\delta+\frac{9}{2}\tau^2)$, and for the gamma distribution it is equal to $a(a+1)(a+2)/b^3$. We are then in the position to compute the skewness of the two components of the mixture approximation to $p(\theta_i|\mu,\sigma^2,\mathbf{y})$, generally given for a random variable X by

$$\frac{\mathrm{E}(X^3)-3\mathrm{E}(X^2)\mathrm{E}(X)+2\mathrm{E}^3(X)}{\{\mathrm{var}(X)\}^{3/2}}.$$

We can therefore obtain the mixing proportion, so that the skewness of the mixture approximation is equal to the skewness of the exact full conditional posterior distribution. If we let β, β_1 and β_2 denote the skewness of the exact full conditional, the log-normal and the gamma distribution respectively, and if ρ is the mixing weight for the log-normal component of the mixture approximation, we only need to solve the equation $\beta = \rho\beta_1 + (1-\rho)\beta_2$ to obtain

$$\rho = \frac{\beta-\beta_2}{\beta_1-\beta_2}. \tag{5.51}$$

We confine ρ to lie within the unit interval by adopting the convention that $\rho = 0$, or $\rho = 1$, whenever the weight produced by (5.51) is respectively negative or greater than 1. Then, to simulate from the resulting log-normal/gamma mixture approximation to the full conditional posterior distribution of θ_i, $i = 1,\ldots,m$,

we sample from a $\mathrm{LN}(\delta, \tau^2)$ distribution with probability ρ and from a $\mathrm{Ga}(a, b)$ with probability $(1-\rho)$, applying the following algorithm.

Step 1. *Generate* $u \sim U(0,1)$

Step 2. *If* $u \leq \rho$, *generate* $\theta_i \sim LN(\delta, \tau^2)$

else, generate $\theta_i \sim Ga(a, b)$.

We note that, following the above procedure, the use of the mixture approximation allows a compromise between a log-normal and a gamma full conditional posterior distribution for the parameters θ_i, $i = 1, \ldots, m$, at each iteration of the Gibbs sampling algorithm.

The remaining question is how to obtain the moments of the exact full conditional posterior distribution of θ_i. As noticed before, the corresponding probability density function in (5.48) implies that we do not have a closed analytical form for these moments, and thus we must employ either an analytical approximation, or a numerical method to derive them. Some of the methods that could be possibly used are listed below:

(a) *Numerical integration.* In theory, this method would provide highly accurate approximations to the exact moments of interest. However, we require two numerical integrations involving the density (5.48) and therefore, taking into cosideration the high computing time cost of the technique, it would be inefficient to combine it with an MCMC scheme.

(b) *Monte Carlo integration.* The problem of estimating the moments of (5.48) was successfully tackled in Chapter 3, employing an importance sampling technique. The method performed remarkably well, provided that the number of Monte Carlo simulations was sufficiently large. Hence, the computer time required for this method to provide accurate results, would again slow down the MCMC algorithm considerably.

(c) *Normal approximations.* Various analytical methods based on normal approximations may be used. We need to account for the non-normal shape of the full conditional posterior distribution, which is due to the form of the likelihood of θ_i. We suggest a discrete approximation to the normal prior distribution of $\gamma_i = \log(\theta_i)$, which relies on matching the first 10 moments of the two distributions.

(d) *Entropy based approximations.* We investigate the use of a method based on

minimising the entropy distance between the exact and an approximating density.

Here, the main concern is acquiring a high degree of accuracy for the resulting approximation, under a relatively small computing time cost. Thus, the computationally intensive nature of the first two methods suggests that they should not be used in this task. On the other hand, we expect methods (c) and (d) to perform well, without being computationally expensive. The accuracy of the proposed methods depends on the values of the observations y_i and the hyperparameters μ and σ^2 involved in (5.48). Due to the iterative nature of the MCMC algorithm, μ and σ^2 may take rather extreme values in a Gibbs sampler cycle, causing substantial inaccuracies in the outcome of either method (c) or (d). However, empirical experimentation suggested that, a combination of the two approaches seems to yield good results. Method (c) performs well when the dispersion of the full conditional posterior distribution of θ_i is small. On the other hand, when the full conditional distribution is highly dispersed, this approach provides poor results, and thus the entropy-based approximation is preferred. The latter leads to accurate posterior moments when the variation of the full conditional distribution under consideration is moderate to high, with the exception of the case when a zero count is observed in the data. In this situation, method (c) seems to perform slightly better, regardless the dispersion of (5.48). In the following sections we describe the two approximating methods.

5.4.4 Discrete approximation to the moments of $p(\theta_i|\mu, \sigma^2, \mathbf{y})$

The full conditional posterior distribution of θ_i given μ and σ^2 depends on the data only through the ith component of $\mathbf{y}$, as shown in (5.48), and therefore can be written as

$$p(\theta_i|\mu, \sigma^2, \mathbf{y}) \propto L(\theta_i|y_i)\ \pi(\theta_i), \qquad i = 1, \ldots, m, \tag{5.52}$$

with

$$L(\theta_i|y_i) \propto \theta_i^{y_i}\ e^{-\theta_i}, \qquad i = 1, \ldots, m \tag{5.53}$$

giving the Poisson likelihood for θ_i, and

$$\pi(\theta_i) \propto \exp\left\{-\frac{1}{2}\ \sigma^{-2}\,(\log\theta_i - \mu)^2 - \log\theta_i\right\}, \qquad i = 1, \ldots, m \tag{5.54}$$

being its LN(μ, σ^2) prior distribution. Letting $\gamma_i = \log(\theta_i)$, the prior distribution of γ_i is $N(\mu, \sigma^2)$.

We will derive a k-point discrete approximation to the normal prior distribution of γ_i, which is easily translated to the corresponding prior of θ_i. We assume that γ_i takes k values

$$\gamma_{ij} = \mu + b_j\sigma, \qquad i = 1, \ldots, m, \; j = 1, \ldots, k, \tag{5.55}$$

with suitably derived probabilities p_j, $j = 1, \ldots, k$. These probabilities and the points b_j, $j = 1, \ldots, k$, define a k-point discrete approximation to the standardised normal distribution $N(0,1)$. Suppose for the moment that we have calculated the appropriate probabilities p_j, $j = 1, \ldots, k$, for a chosen grid of k γ_{ij} points, a problem to which we will return later in this subsection. Then, (5.55) means that, given μ and σ^2, each θ_i takes the following, denoted as d_j, values

$$\theta_{ij} = e^{\mu + b_j\sigma} = d_j, \tag{5.56}$$

with approximate prior probabilities p_j, $j = 1, \ldots, k$. We therefore have a k-point discrete approximation to the prior distribution of the Poisson means θ_i.

If we now let q_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, k$ denote the posterior probabilities that $\theta_{ij} = d_j$, then Bayes' theorem implies that these posterior probabilities will be approximately given by

$$q_{ij} \propto L(\theta_i|y_i)\; \pi^*(\theta_i), \qquad i = 1, \ldots, m, \quad j = 1, \ldots, k,$$

where $L(\theta_i|y_i)$ is the Poisson likelihood in (5.53) and $\pi^*(\theta_i)$ denotes the discrete approximation to the prior distribution of θ_i. Hence, given μ and σ^2, we approximately have that each Poisson mean takes values $\theta_{ij} = d_j$ with posterior probabilities given by

$$q_{ij} = \frac{d_j^{y_i} e^{-d_j} p_j}{\sum_{l=1}^{k} d_l^{y_i} e^{-d_l} p_l}, \tag{5.57}$$

for $i = 1, \ldots, m$, $j = 1, \ldots, k$. The above equation defines a k-point discrete approximation to the full conditional distribution $p(\theta_i|\mu, \sigma^2, \mathbf{y})$. Employing the points d_j from (5.56) and the probabilities q_{ij}, $j = 1, \ldots, k$ from (5.57), we can now obtain approximations to the moments of the full conditional distribution of θ_i. The rth order moment may be computed as

$$\mathrm{E}(\theta_i^r|\mu, \sigma^2, \mathbf{y}) \doteq \sum_{j=1}^{k} d_j^r q_{ij}. \tag{5.58}$$

Discretisation of standard normal distribution

We can now return to the question of how to derive the k-point discrete approximation to the standardised normal distribution. This is done by matching the

Table 5.4: *Moments of even order for the normal $N(0,1)$ distribution.*

r	1	2	3	4	5
$\mathrm{E}(Z^{2r})$	1	3	15	105	945

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 equations (e.g., Stuart and Ord, 1994)

$$\mathrm{E}\left(Z^{2r}\right) = \frac{(2r)!}{2^r r!}, \qquad r = 1, \ldots, 5, \tag{5.59}$$

giving the moments presented in Table 5.4.

For a symmetrical distribution of a discrete random variable X, evaluated at the points $b_1, b_2, \ldots, b_k$ with probabilities $p_1, p_2, \ldots, p_k$, the corresponding moments of even order are given by

$$\mathrm{E}\left(X^{2r}\right) = \sum_{j=1}^{k} b_j^{2r} p_j, \qquad r = 1, \ldots, 5. \tag{5.60}$$

We take the points b_j to be equally spaced on a suitably selected grid, with the distance between two successive points equal to a fixed value b. Then, due to the symmetry of the normal $N(0,1)$ distribution around the origin we notice that we only need to specify the probabilities $p_0, p_1, p_2, \ldots, p_{\frac{k-1}{2}}$, which correspond to the points lying on the non-negative part the x-axis, that is the points $0, b, 2b, \ldots, \frac{k-1}{2}b$. Clearly, the same probabilities correspond to the the equivalent points on the negative axis. This implies that (5.60) can also be written as

$$\begin{aligned} \mathrm{E}\left(X^2\right) &= 2 \sum_{j=0}^{\frac{k-1}{2}} (jb)^2 p_j \\ &\vdots \\ \mathrm{E}\left(X^{10}\right) &= 2 \sum_{j=0}^{\frac{k-1}{2}} (jb)^{10} p_j \end{aligned}$$

leading to the general form

$$\mathrm{E}\left(X^{2r}\right) = 2 \sum_{j=0}^{\frac{k-1}{2}} (jb)^{2r} p_j, \qquad r = 1, \ldots, 5. \tag{5.61}$$

Thus, using (5.59) and (5.61) we can match the first 10 moments of the normal

$N(0,1)$ and the approximating distribution by solving the equations

$$\frac{(2r)!}{2^r r!} = 2 \sum_{j=0}^{\frac{k-1}{2}} (jb)^{2r} p_j, \qquad r = 1, \ldots, 5, \tag{5.62}$$

to obtain the probabilities $p_1, p_2, \ldots, p_{\frac{k-1}{2}}$ and $p_0 = 1 - 2\sum_{j=1}^{\frac{k-1}{2}} p_j$. Choosing to match the first 10 moments of the two distributions gives the five equations (5.61), each of which produces the probabilities of two points for the discrete distribution. These, together with the probability of the point at the origin, provide a discrete approximation evaluated at a maximum of $k = 11$ points. This choice, combined with fixing the distance between successive points to $b = 0.8$, seemed to yield a good approximation of this kind to the full conditional distribution of θ_i. The distance b does not have to be fixed. In theory, we can calculate it as part of the set of equations (5.62), by including an extra equation for $r = 6$ and keeping the number of points to $k = 11$. However, for all practical purposes, we found it easier to assign a fixed value to b.

As mentioned before, the method that we have presented performs outstandingly when the variation of the full conditional distribution of θ_i is relatively small. Nevertheless, it does not seem to exhibit the desired accuracy uniformly for all μ, σ^2 and y values. Since these values may take a large number of extreme combinations, due to the iterative nature of the Gibbs algorithm, which may in turn lead to a highly dispersed full conditional distribution, we suggest this method to be used alternately with an entropy-based approach described in the following subsection.

5.4.5 Entropy based approximation to the moments of $p(\theta_i|\mu, \sigma^2, \mathbf{y})$

As before, the aim is to derive approximations to the moments of the full conditional distribution of θ_i, that is expectations of the form $\mathrm{E}(\theta_i^r|\mu, \sigma^2, \mathbf{y})$. Employing the reparametrisation $\gamma_i = \log(\theta_i)$, these are given by

$$\begin{aligned} \mathrm{E}\left(\theta_i^r|\mu, \sigma^2, \mathbf{y}\right) &= \int_{-\infty}^{\infty} e^{r\gamma_i}\, p(\gamma_i|\mu, \sigma^2, \mathbf{y})\, d\gamma_i \\ &= \int_{-\infty}^{\infty} e^{r\gamma_i}\, \frac{p(y_i|\gamma_i)\, \pi(\gamma_i|\mu, \sigma^2)}{p(y_i)}\, d\gamma_i, \end{aligned}$$

for $i = 1, \ldots, m$, where $p(y_i|\gamma_i)$ is the sampling density, $\pi(\gamma_i|\mu, \sigma^2)$ is the prior distribution of γ_i and $p(y_i)$ denotes the marginal density of the observation y_i. Then according to the model specification, using the Poisson sampling density we

have that

$$\mathrm{E}\left(\theta_i^r|\mu,\sigma^2,\mathbf{y}\right) = \int_{-\infty}^{\infty} \frac{\exp\left\{\gamma_i(y_i+r)-e^{\gamma_i}\right\}}{p(y_i)\ y_i!}\ \pi(\gamma_i|\mu,\sigma^2)\ d\gamma_i$$

$$= \frac{(y_i+r)!}{p(y_i)\ y_i!} \int_{-\infty}^{\infty} p(y_i+r|\gamma_i)\ \pi(\gamma_i|\mu,\sigma^2)\ d\gamma_i$$

$$= \frac{(y_i+r)!}{y_i!}\ \frac{p(y_i+r)}{p(y_i)}\,. \tag{5.63}$$

The above demonstrates that we can obtain the approximate conditional posterior moments $\mathrm{E}(\theta_i^r|\mu,\sigma^2,\mathbf{y})$, only by deriving an approximation to the marginal density of the data.

An entropy approximation based on an iterative procedure

We consider the problem of the approximation of the joint marginal density of the data in a multivariate setting, which is equivalent to that of model (4.1a). We let $\mathbf{y} = (y_1,\ldots,y_m)^{\mathrm{T}}$ denote the data vector and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_m)^{\mathrm{T}}$ the vector consisting of the m Poisson means. We also take $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_m)^{\mathrm{T}}$ to be the m-dimensional vector with ith element equal to $\gamma_i = \log(\theta_i)$. Then, the random variables $Y_1,\ldots,Y_m$, conditional on $\theta_1,\ldots,\theta_m$, are distributed according to independent Poisson(θ_i) distributions. In the prior assessment, we assume that the vector $\boldsymbol{\gamma}$ has an m-dimensional normal distribution with mean vector denoted by $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_m)^{\mathrm{T}}$ and covariance matrix $\mathbf{C}$, which is an $m \times m$ symmetrical matrix with its (i,j) element denoted by c_{ij}. Both $\boldsymbol{\mu}$ and and $\mathbf{C}$ are assumed known, which will be the case in the context of the Gibbs sampling procedure. The model can be written as

$$\begin{aligned} Y_i \mid \gamma_i &\sim \text{Poisson}\left(e^{\gamma_i}\right), \qquad i = 1,\ldots,m \\ \boldsymbol{\gamma} \mid \boldsymbol{\mu},\mathbf{C} &\sim N_m\left(\boldsymbol{\mu},\mathbf{C}\right). \end{aligned} \tag{5.64}$$

For the remaining of this section we suppress the dependency on $\boldsymbol{\mu}$ and and $\mathbf{C}$ (or μ and σ^2) in the notation, wherever this does not affect our presentation. The joint density of $\mathbf{y}$ and $\boldsymbol{\gamma}$ is given by

$$p(\mathbf{y},\boldsymbol{\gamma}) = p(\mathbf{y}|\boldsymbol{\gamma})\ \pi(\boldsymbol{\gamma}),$$

where $p(\mathbf{y}|\boldsymbol{\gamma})$ is the Poisson conditional density and $\pi(\boldsymbol{\gamma})$ is the normal prior distribution of $\boldsymbol{\gamma}$. Thus, from model (5.64), the above gives

$$p(\mathbf{y},\boldsymbol{\gamma}) = (2\pi)^{-\frac{m}{2}} \left\{\prod_{i=1}^{m}(y_i!)\right\}^{-1} |\mathbf{C}|^{-\frac{1}{2}} \exp\left\{\sum_{i=1}^{m} y_i\gamma_i - \sum_{i=1}^{m} e^{\gamma_i} - \frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{C}^{-1}(\boldsymbol{\gamma}-\boldsymbol{\mu})\right\}. \quad (5.65)$$

We consider approximations to (5.65) of the form

$$p^*(\mathbf{y},\boldsymbol{\gamma}) = p^*(\mathbf{y})\, p^*(\boldsymbol{\gamma}|\mathbf{y}), \quad (5.66)$$

where $p^*(\boldsymbol{\gamma}|\mathbf{y})$ is an m-dimensional normal $N_m(\boldsymbol{\alpha}, \mathbf{G})$ density, that is

$$p^*(\boldsymbol{\gamma}|\mathbf{y}) = (2\pi)^{-\frac{m}{2}} |\mathbf{G}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\alpha})^{\mathrm{T}}\mathbf{G}^{-1}(\boldsymbol{\gamma}-\boldsymbol{\alpha})\right\} \quad (5.67)$$

and $p^*(\mathbf{y})$ approximates the marginal density of the data. We will estimate the parameters of the $N_m(\boldsymbol{\alpha}, \mathbf{G})$ distribution by minimising the entropy distance between $p(\mathbf{y},\boldsymbol{\gamma})$ and $p^*(\mathbf{y},\boldsymbol{\gamma})$.

Definition 5.6. *We let X be a random variable with density $f(x)$. Then the entropy associated with the density $f(\cdot)$, denoted by I_f, is defined as*

$$I_f = -\int \log f(x) f(x)\, dx = E\{-\log f(X)\},$$

and the entropy distance between some other density function $h(X)$ and $f(X)$ is defined as

$$\mathcal{D}I = I_h - I_f = \int \{\log f(x) - \log h(x)\} f(x)\, dx = E\left\{\log \frac{f(X)}{h(X)}\right\}, \quad (5.68)$$

where the expectation is with respect to $f(X)$.

Applications of the entropy idea in subjects including information measures and construction of general probability models or prior probability settings, appear in Lindley (1956) and Rosenkrantz (1989).

Let us now denote the entropy distance between the approximating density $p^*(\mathbf{y},\boldsymbol{\gamma})$ and the exact joint density of $\mathbf{y}$ and $\boldsymbol{\gamma}$ by $\mathcal{D}I(\boldsymbol{\alpha}, \mathbf{G})$, to express it as a function of the parameters under estimation. Then, this can be written as

$$\mathcal{D}I(\boldsymbol{\alpha}, \mathbf{G}) = \mathrm{E}\left\{\log \frac{p^*(\mathbf{y},\boldsymbol{\gamma})}{p(\mathbf{y},\boldsymbol{\gamma})}\right\}, \quad (5.69)$$

where the expectation is taken with respect to the approximate distribution of $\boldsymbol{\gamma}$ given $\mathbf{y}$, that is the m-dimensional normal $N_m(\boldsymbol{\alpha}, \mathbf{G})$ distribution. From equations (5.65), (5.66) and (5.67), it follows that the logarithm of the ratio involved in (5.69) is given by

$$\log \frac{p^*(\mathbf{y}, \boldsymbol{\gamma})}{p(\mathbf{y}, \boldsymbol{\gamma})} = \log p^*(\mathbf{y}) + \log \left\{ \prod_{i=1}^{m} (y_i!) \right\} + \frac{1}{2} \log|\mathbf{C}| - \frac{1}{2} \log|\mathbf{G}|$$
$$- \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{y} + \sum_{i=1}^{m} e^{\gamma_i} - \frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\alpha})^{\mathrm{T}} \mathbf{G}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\alpha}) + \frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{C}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\mu}). \quad (5.70)$$

We notice that the last two terms of the above equation, can be written as

$$- \frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\alpha})^{\mathrm{T}} \mathbf{G}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\alpha}) + \frac{1}{2} (\boldsymbol{\gamma} - \boldsymbol{\mu})^{\mathrm{T}} \mathbf{C}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\mu}) =$$
$$- \frac{1}{2} \operatorname{tr} \left\{ \mathbf{G}^{-1} (\boldsymbol{\gamma} - \boldsymbol{\alpha})(\boldsymbol{\gamma} - \boldsymbol{\alpha})^{\mathrm{T}} - \mathbf{C}^{-1} \left(\boldsymbol{\gamma}\boldsymbol{\gamma}^{\mathrm{T}} - 2\boldsymbol{\mu}\boldsymbol{\gamma}^{\mathrm{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \right) \right\},$$

and therefore, taking the expectation of expression (5.70) with respect to the $N_m(\boldsymbol{\alpha}, \mathbf{G})$ density for $\boldsymbol{\gamma}$, we obtain

$$\mathcal{D}I(\boldsymbol{\alpha}, \mathbf{G}) = \log p^*(\mathbf{y}) + \log \left\{ \prod_{i=1}^{m} (y_i!) \right\} + \frac{1}{2} \log|\mathbf{C}| - \frac{1}{2} \log|\mathbf{G}| - \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{y} + \sum_{i=1}^{m} e^{\alpha_i + \frac{1}{2} g_{ii}}$$
$$- \frac{1}{2} \operatorname{tr} \left\{ \mathbf{G}^{-1}\mathbf{G} - \mathbf{C}^{-1} \left(\mathbf{G} + \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}} \right) + 2\mathbf{C}^{-1}\boldsymbol{\mu}\boldsymbol{\alpha}^{\mathrm{T}} - \mathbf{C}^{-1}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} \right\},$$

which gives

$$\mathcal{D}I(\boldsymbol{\alpha}, \mathbf{G}) = \log p^*(\mathbf{y}) + \log \left\{ \prod_{i=1}^{m} (y_i!) \right\} + \frac{1}{2} \log|\mathbf{C}| - \frac{1}{2} \log|\mathbf{G}| - \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{y} + \sum_{i=1}^{m} e^{\alpha_i + \frac{1}{2} g_{ii}}$$
$$- \frac{m}{2} + \frac{1}{2} \operatorname{tr} \left\{ \mathbf{C}^{-1} (\mathbf{G} + \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}) \right\} - \boldsymbol{\alpha}^{\mathrm{T}} \mathbf{C}^{-1} \boldsymbol{\mu} + \frac{1}{2} \boldsymbol{\mu}^{\mathrm{T}} \mathbf{C}^{-1} \boldsymbol{\mu}, \quad (5.71)$$

where α_i is the ith component of the mean vector $\boldsymbol{\alpha}$ and g_{ii} the ith diagonal element of the covariance matrix $\mathbf{G}$. We can now minimise (5.71) for $\boldsymbol{\alpha}$ and $\mathbf{G}$. We first differentiate the function $\mathcal{D}I(\boldsymbol{\alpha}, \mathbf{G})$ with respect to the vector $\boldsymbol{\alpha}$, using the results in Appendix C, to obtain

$$\frac{\partial \mathcal{D}I(\boldsymbol{\alpha}, \mathbf{G})}{\partial \boldsymbol{\alpha}} = -\mathbf{y} + \exp\left(\boldsymbol{\alpha} + \frac{1}{2} \mathbf{g} \right) + \mathbf{C}^{-1} (\boldsymbol{\alpha} - \boldsymbol{\mu}),$$

where $\mathbf{g}$ is the m-dimensional vector consisting of the diagonal elements of the covariance matrix $\mathbf{G}$. Setting the above expression equal to zero, we obtain the equation

$$\exp\left(\boldsymbol{\alpha} + \frac{1}{2} \mathbf{g} \right) = \mathbf{y} - \mathbf{C}^{-1} (\boldsymbol{\alpha} - \boldsymbol{\mu}). \quad (5.72)$$

Then, differentiation with respect to matrix $\mathbf{G}$ (again using the results in Appendix C), gives

$$\frac{\partial \mathcal{D}I(\boldsymbol{\alpha}, \mathbf{G})}{\partial \mathbf{G}} = \frac{1}{2}\left(-\mathbf{G}^{-1} + \mathbf{F} + \mathbf{C}^{-1}\right),$$

with $\mathbf{F}$ denoting the diagonal matrix with elements $e^{\alpha_i + \frac{1}{2}g_{ii}}$, $i = 1, \ldots, m$, on its diagonal. Equating the above derivative to zero, will give

$$\mathbf{G}^{-1} = \mathbf{F} + \mathbf{C}^{-1}. \tag{5.73}$$

Clearly, the solution to equations (5.72) and (5.73) for $\boldsymbol{\alpha}$ and $\mathbf{G}$ requires an iterative procedure. However, for any given $\boldsymbol{\alpha}$ and $\mathbf{G}$, the approximating density $p^*(\mathbf{y}, \boldsymbol{\gamma})$ and the exact joint density $p(\mathbf{y}, \boldsymbol{\gamma})$, are identical when (5.71) is equal to zero. Therefore, if $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\mathbf{G}}$ satisfy the equations (5.72) and (5.73), and their corresponding elements are denoted by $\tilde{\alpha}_i$ and $\tilde{g}_{ij}$, then the marginal density of $\mathbf{y}$ which sets the entropy distance (5.71) equal to zero, is given by

$$p^*(\mathbf{y}) = |\tilde{\mathbf{G}}|^{\frac{1}{2}}\ |\mathbf{C}|^{-\frac{1}{2}} \left\{\prod_{i=1}^{m}(y_i!)\right\}^{-1} \exp\left\{\frac{m}{2} + \tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{y} - \sum_{i=1}^{m} e^{\tilde{\alpha}_i + \frac{1}{2}\tilde{g}_{ii}}\right\}$$
$$\exp\left[-\frac{1}{2}\ \mathrm{tr}\left\{\mathbf{C}^{-1}\left\{\tilde{\mathbf{G}} + (\tilde{\boldsymbol{\alpha}} - \boldsymbol{\mu})(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\mu})^{\mathrm{T}}\right\}\right\}\right]. \tag{5.74}$$

If we now consider the same model in a component-wise form, that is model (4.1a), we notice that we seek approximations to the marginal density of the observation y_i, denoted by $p(y_i)$. Then, the multivariate version (5.74), implies that the density $p^*(y_i)$ which minimises the entropy distance between the exact joint density $p(y_i, \gamma_i)$ and an approximation of the form

$$p^*(y_i, \gamma_i) = p^*(y_i)\ p^*(\gamma_i | y_i),$$

where $p^*(\gamma_i|y_i)$ is the probability density function of a $N(\alpha_i, g_{ii})$ distribution, will be given by

$$p^*(y_i) = \tilde{g}_{ii}^{\frac{1}{2}}\ (\sigma^2)^{-\frac{1}{2}}\ (y_i!)^{-1} \exp\left(\frac{1}{2} + \tilde{\alpha}_i y_i - e^{\tilde{\alpha}_i + \frac{1}{2}\tilde{g}_{ii}}\right)$$
$$\exp\left[-\frac{1}{2}\ \sigma^{-2}\left\{\tilde{g}_{ii} + (\tilde{\alpha}_i - \mu)^2\right\}\right], \tag{5.75}$$

where $\tilde{\alpha}_i$ and $\tilde{g}_{ii}$ are the iterative solutions to the set of equations

$$\begin{aligned} e^{\alpha_i + \frac{1}{2}g_{ii}} &= y_i - \sigma^{-2}\ (\alpha_i - \mu) \\ g_{ii}^{-1} &= e^{\alpha_i + \frac{1}{2}g_{ii}} + \sigma^{-2} \end{aligned} \tag{5.76}$$

in complete accordance to (5.72) and (5.73). Substitution of the approximation (5.75) for $p(y_i)$, in (5.63) will yield the desired approximation to the conditional posterior moments $\mathrm{E}(\theta_i|\mu, \sigma^2, \mathbf{y})$. However, we can avoid the iterative nature of the solutions for $\tilde{\alpha}_i$ and $\tilde{g}_{ii}$ used in (5.75), by deriving an algebraically explicit approximation, which is presented in the next paragraph. Employing an analytical solution, rather than having to take iterative steps in order to reach an explicit form for $p^*(y_i)$, might prove substantially useful, since we intend to incorporate the procedure for obtaining the approximations within each iteration of the MCMC algorithm. Nevertheless, if the iterative solution is preferred, the explicit solution may be used as a good starting point for the computational procedure.

An analytically explicit approximation

In the preceding analysis we employed a normal approximation to the posterior distribution γ_i, and then obtained an iterative solution for α_i and g_{ii}, the respective mean and variance of the normal density. We will now derive an algebraically explicit approximation to $p(\gamma_i|y_i)$, which may be used directly for the calculation of $p^*(y_i)$ in (5.75). We follow the same entropy-based procedure of the preceding section.

The posterior distribution of γ_i can be expressed as

$$p(\gamma_i|y_i) = \frac{L(\gamma_i|y_i)\ \pi(\gamma_i)}{p(y_i)}, \tag{5.77}$$

where $\pi(\gamma_i)$ is the prior distribution of γ_i and $L(\gamma_i|y_i)$ is the likelihood given by

$$L(\gamma_i|y_i) = \frac{\exp\left(\gamma_i y_i - e^{\gamma_i}\right)}{y_i!}.$$

Then, for any positive real number t, the likelihood for γ_i can be rewritten as

$$\begin{aligned} L(\gamma_i|y_i) &= \frac{e^{-t\gamma_i}\ \Gamma(y_i+t)}{y_i!}\ \frac{\exp\left\{\gamma_i(y_i+t) - e^{\gamma_i}\right\}}{\Gamma(y_i+t)} \\ &= \frac{e^{-t\gamma_i}\ \Gamma(y_i+t)}{\Gamma(y_i+1)}\ L(\gamma_i|y_i+t), \end{aligned} \tag{5.78}$$

where

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

gives the probability density function of the logarithm of a $\mathrm{Ga}(y_i+t, 1)$ random variable. Hence, we suggest that it would be reasonable to approximate $L(\gamma_i|y_i+t)$ with the probability density function of a normal $N(\delta_i, v_i)$ distribution, denoted by $L^*(\gamma_i|y_i+t)$. We will specify the mean δ_i and the variance v_i of the normal

approximation, in such a manner that they minimise the entropy distance between the density of the $N(\delta_i, v_i)$ distribution and the $L(\gamma_i|y_i + t)$ density. If we let $\mathcal{DI}(\delta_i, v_i)$ denote this entropy distance, we have that

$$\mathcal{DI}(\delta_i, v_i) = \mathrm{E}\left\{\log \frac{L^*(\gamma_i|y_i + t)}{L(\gamma_i|y_i + t)}\right\},$$

where the expectation corresponds to the $N(\delta_i, v_i)$ distribution for γ_i. Thus, the above gives

$$\begin{aligned}\mathcal{DI}(\delta_i, v_i) &= \mathrm{E}\left[\log\left\{(2\pi)^{-\frac{1}{2}} v_i^{-\frac{1}{2}} \{\Gamma(y_i + t)\}^{-1}\right\} - \frac{1}{2}\, v_i^{-1}(\gamma_i - \delta_i)^2 - \gamma_i(y_i + t) + e^{\gamma_i}\right] \\ &= -\frac{1}{2}\log(2\pi) - \log\{\Gamma(y_i + t)\} - \frac{1}{2}\log v_i - \frac{1}{2} - \delta_i(y_i + t) + e^{\delta_i + \frac{1}{2}v_i}. \qquad (5.79)\end{aligned}$$

The entropy distance (5.79) is minimised for δ_i and v_i satisfying the equations $\frac{\partial \mathcal{DI}(\delta_i, v_i)}{\partial \delta_i} = 0$ and $\frac{\partial \mathcal{DI}(\delta_i, v_i)}{\partial v_i} = 0$, which will give

$$-(y_i + t) + e^{\delta_i + \frac{1}{2}v_i} = 0 \qquad \text{and} \qquad -\frac{1}{2}v_i^{-1} + \frac{1}{2}e^{\delta_i + \frac{1}{2}v_i} = 0.$$

From these two equations it follows that

$$\begin{aligned}\delta_i &= \log(y_i + t) - \frac{1}{2}\,(y_i + t)^{-1} \\ v_i &= (y_i + t)^{-1}.\end{aligned} \qquad (5.80)$$

Now, from (5.78), we notice that the likelihood function for γ_i is proportional to $e^{-t\gamma_i}\, L(\gamma_i|y_i + t)$, and if we replace $L(\gamma_i|y_i + t)$ by its normal $N(\delta_i, v_i)$ approximation with parameters given in (5.80), we obtain

$$\begin{aligned}L(\gamma_i|y_i) &\propto v_i^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\, v_i^{-1}(\gamma_i - \delta_i)^2 - t\gamma_i\right\} \\ &\propto \exp\left[-\frac{1}{2}\, v_i^{-1}\left\{\gamma_i^2 + \delta_i^2 - 2\gamma_i\delta_i + 2t\gamma_i v_i\right\}\right] \\ &\propto \exp\left[-\frac{1}{2}\, v_i^{-1}\left\{\gamma_i - (\delta_i - tv_i)\right\}^2\right].\end{aligned}$$

Clearly, the above expression implies that the likelihood $L(\gamma_i|y_i)$ is proportional to the probability density function of a normal $N(l_i, v_i)$ distribution, with mean

$$l_i = \delta_i - tv_i = \log(y_i + t) - \left(\frac{1}{2} + t\right)(y_i + t)^{-1}$$

and variance v_i given in (5.80). If we take $t = \frac{1}{2}$, the mean and the variance of the normal $N(l_i, v_i)$ approximation to the γ_i likelihood $L(\gamma_i|y_i)$ become

$$\begin{aligned}l_i &= \log\left(y_i + \frac{1}{2}\right) - \left(y_i + \frac{1}{2}\right)^{-1} \\ v_i &= \left(y_i + \frac{1}{2}\right)^{-1}.\end{aligned} \qquad (5.81)$$

This can now be combined with the normal $N(\mu, \sigma^2)$ prior distribution for γ_i, to give an approximation to the posterior distribution of γ_i. As before, we let $p^*(\gamma_i|y_i)$ denote this approximation. The conjugacy of the normal $N(l_i, v_i)$ likelihood and the normal $N(\mu, \sigma^2)$ prior distribution implies that, $p^*(\gamma_i|y_i)$ will also be a normal density, located at

$$\tilde{\alpha}_i = \frac{v_i^{-1} l_i + \sigma^{-2}\mu}{v_i^{-1} + \sigma^{-2}} \tag{5.82}$$

and with variance

$$\tilde{g}_{ii} = \left(v_i^{-1} + \sigma^{-2}\right)^{-1}. \tag{5.83}$$

We have therefore obtained analytically explicit expressions for $\tilde{\alpha}_i$ and $\tilde{g}_{ii}$ which can be used directly in (5.75) to yield an approximation to the marginal distribution $p(y_i)$, which in turn will provide an estimation of the conditional posterior moments $\mathrm{E}\left(\theta_i^r|\mu, \sigma^2, \mathbf{y}\right)$ in (5.63).

We illustrate the performance of the two approximating methods in Figures (5.2) and (5.3). We will refer to the method based on the discrete approximation to the posterior moments of θ_i as the 'discretisation' method, and to the entropy-based technique for the computation of the posterior moments as the 'entropy' method. The graphs compare the exact full conditional distribution of θ_i, which was calculated using numerical integration, with the approximate distributions. In all graphs, the solid line shows the exact density, while the dotted and dashed lines represent the densities obtained with the discretisation and the entropy method respectively. Figure 5.2 exhibits the performance of the two methods, when a zero count is observed. We used 4 different combinations of the hyperparameters μ and σ^2, leading to increasing values of the coefficient of variation for θ_i, ranging from 0.3 to 1.9. Figure 5.3 displays the case when $y = 5$, for the same values of the variation coefficient. Both figures demonstrate the outstanding performance of the discretisation method, when the variation of the distribution is relatively low. As shown in Figures 5.2(a) and 5.3(a), when the coefficient of variation is as low as 0.3, the approximate density based on the discretisation method is indistinguishable from the exact density, irrespectively from the value of y. Figure 5.2 suggests that when $y = 0$, this method seems to be slightly better than the entropy method, also for the remaining values of the variation coefficient. For a zero count, the entropy method seems to produce an approximation which is somewhat dislocated when compared to the exact distribution. On the other hand, with larger observations, as Figure 5.3 illustrates for $y = 5$, the entropy method produces very accurate approximations, whenever the variation of the distribution is not considerably small.

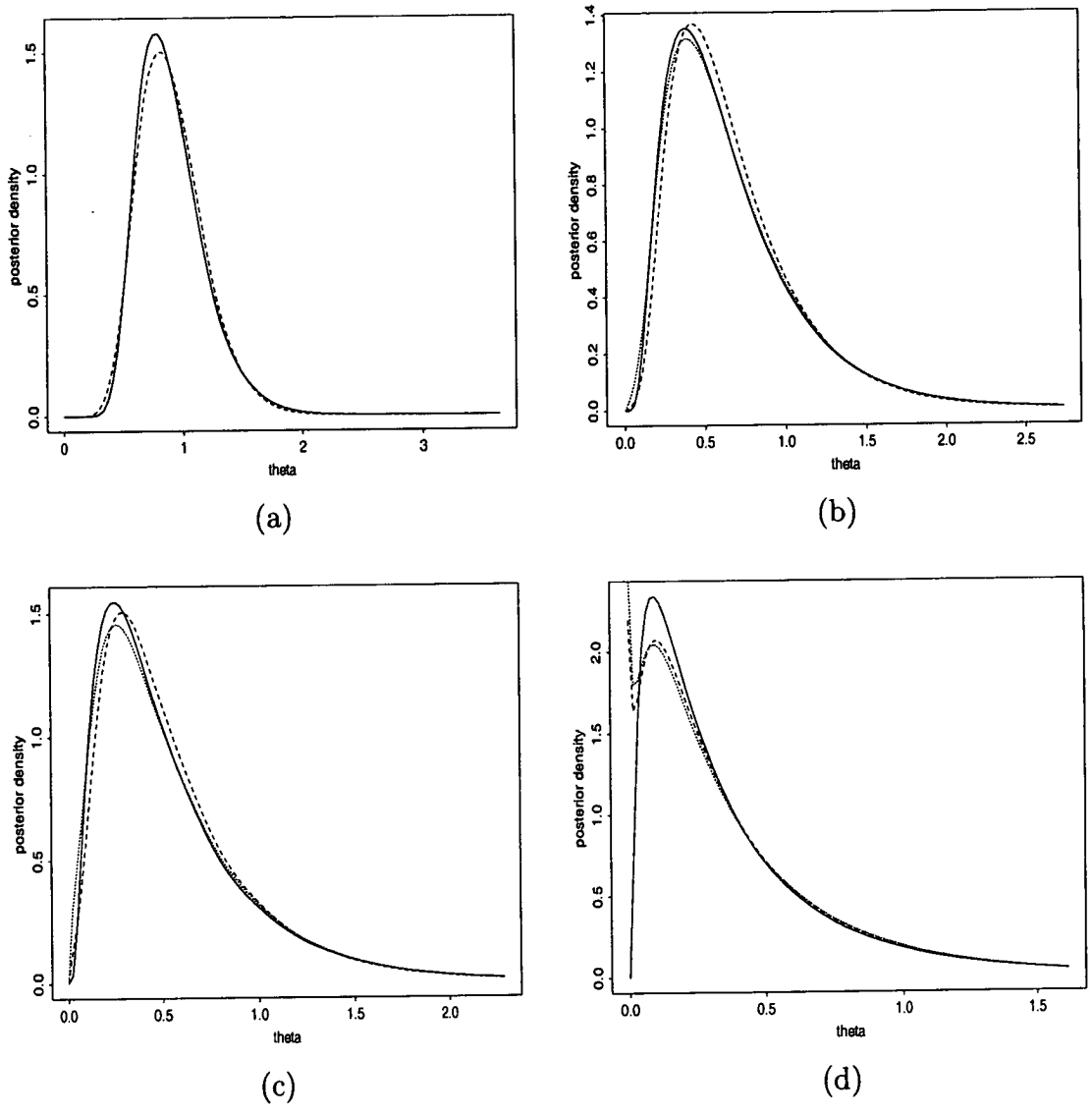

Figure 5.2: *Mixture approximation to* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$, *when the discretisation method (dotted line) or the entropy method (dashed line) is used for the computation of the posterior moments. The solid line corresponds to the exact density. The observation value is* $y = 0$ *and the coefficient of variation is equal to: (a)* 0.3*; (b)* 0.8*; (c)* 1.1*; (d)* 1.9.

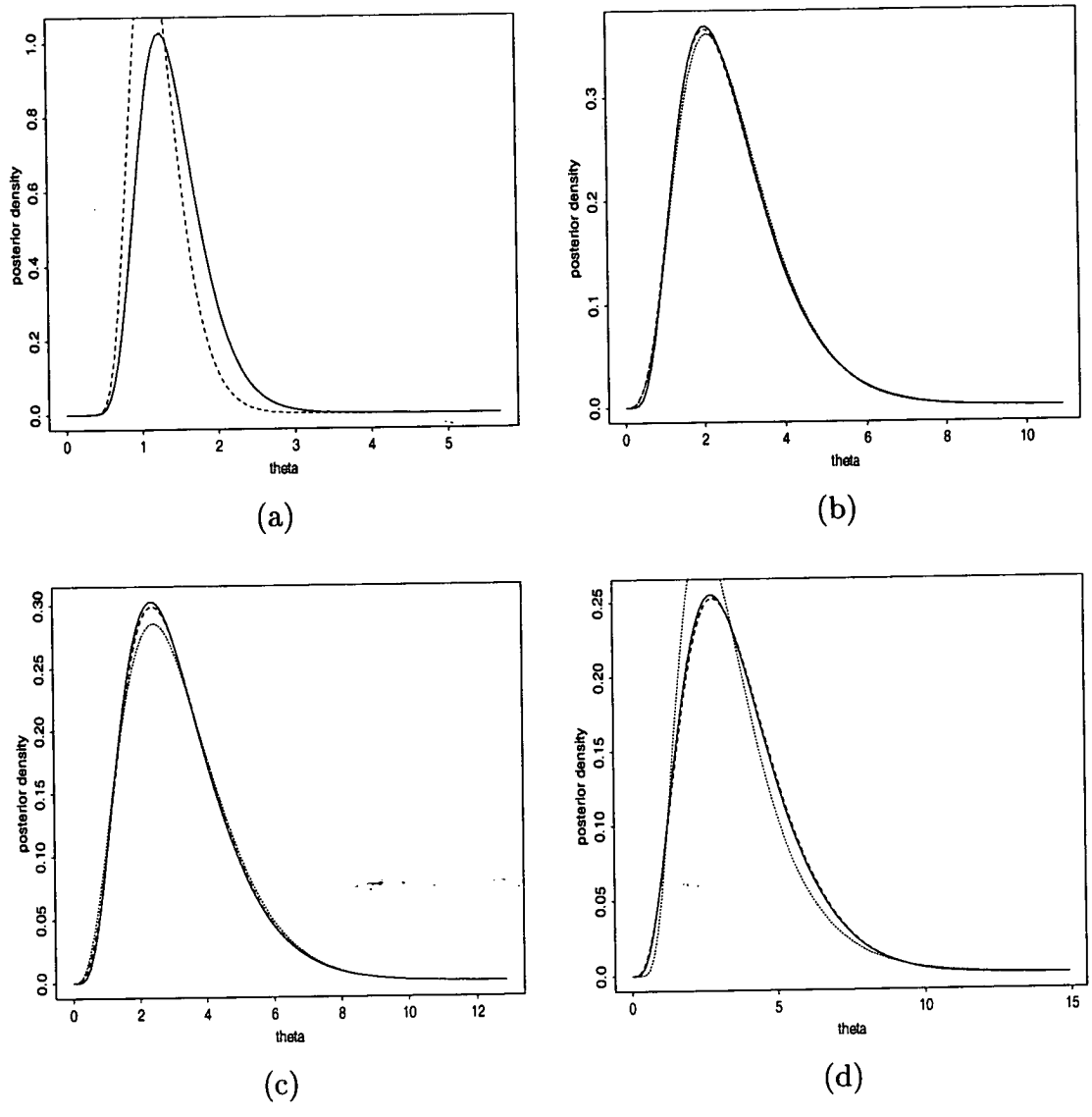

Figure 5.3: *Mixture approximation to* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$, *when the discretisation method (dotted line) or the entropy method (dashed line) is used for the computation of the posterior moments. The solid line corresponds to the exact density. The observation value is* $y = 5$ *and the coefficient of variation is equal to: (a)* 0.3*; (b)* 0.8*; (c)* 1.1*; (d)* 1.9.

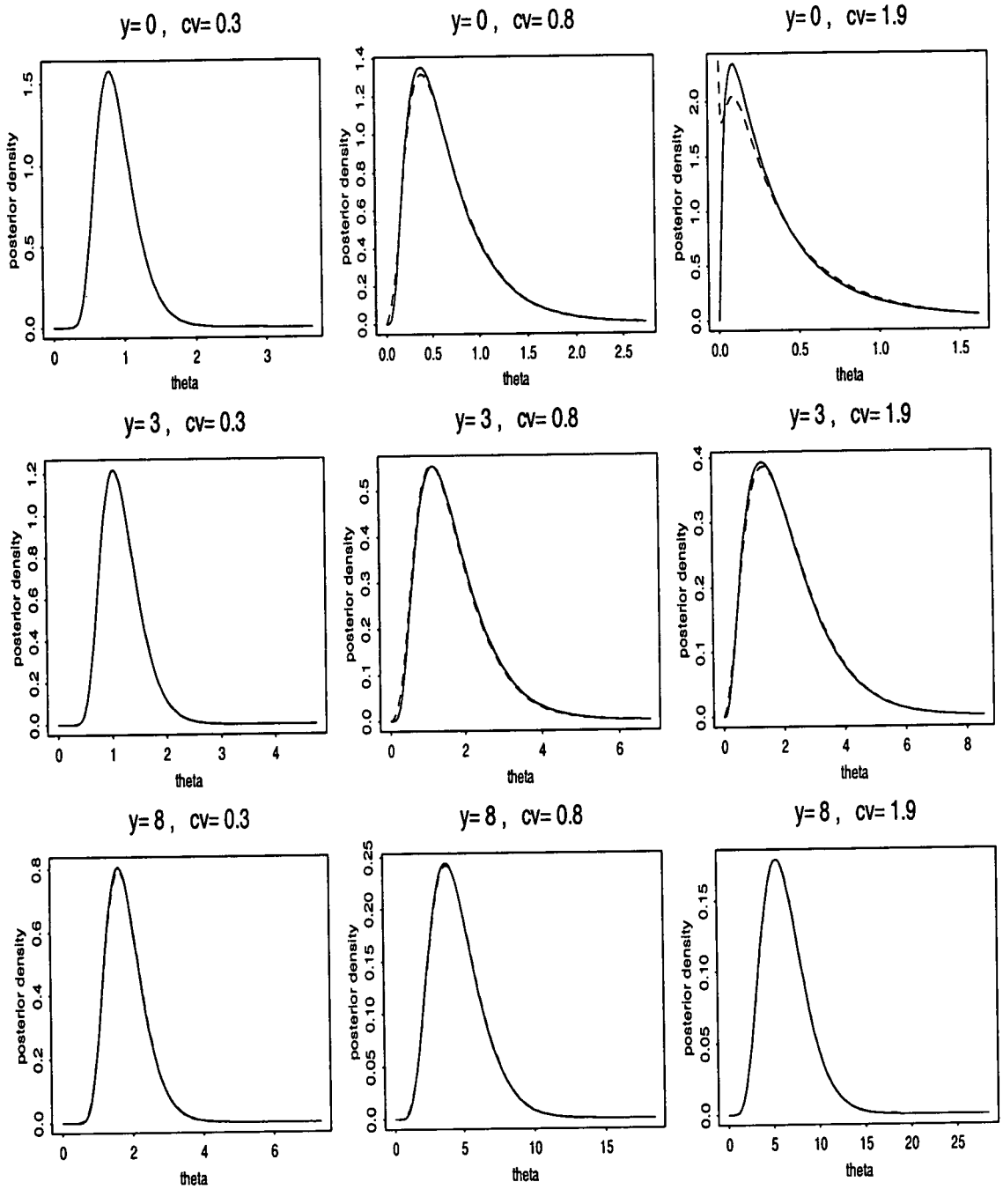

Figure 5.4: *Mixture approximation (dashed line) to* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$ *for different data and variation coefficient values, when a combination of the discretisation and the entropy method is used for the computation of the posterior moments. The solid line corresponds to the exact density.*

These findings indicate that a suitable combination of the two methods would provide a better approximation to the full conditional posterior distribution of θ_i, than any of the described techniques alone. This is also empirically verified from the examples in the following section. The combination strategy that we adopt is that the discretisation method should be preferred when $y = 0$, or when the variation coefficient of θ_i is relatively small, say less than 0.5, whereas the entropy method should be applied in all remaining cases. The approximation produced when we follow this strategy is shown in Figure 5.4, where the exact and approximate densities are plotted for different data and variation coefficient values.

5.4.6 Example: Audit data

We now illustrate the Gibbs sampling method based on the approximations presented in Subsections 5.4.4 and 5.4.5, reanalysing the audit data set which was also considered in Section 3.6 and Subsection 5.2.3. The data consist of the number of errors in audit samples of 9 different accounts and are given in Table 3.1. The model specification is identical to that of Subsection 5.2.3, as again we consider both the uniform $U(0, \infty)$ and the Inv-$\chi^2(10, 0.45)$ hyperprior distributions for σ^2, at the second stage of the prior setting.

The Gibbs sampler will involve simulation from the full conditional posterior distributions of μ, σ^2 and θ_i, $i = 1, \ldots, m$. Following our presentation in Subsection 5.4.2, we will draw the μ values from a $N\left(\bar{\gamma}, \frac{\sigma^2}{m}\right)$ distribution, and the σ^2 variates will be generated such that $\frac{\sum_{i=1}^{m}(\gamma_i-\mu)^2}{\sigma^2}$ has a χ^2_{m-2} distribution, or $\frac{\nu\lambda+\sum_{i=1}^{m}(\gamma_i-\mu)^2}{\sigma^2}$ follows a $\chi^2_{\nu+m}$ distribution, depending on which of the two second stage prior specifications is assumed. To simulate the θ_i values we will employ the log-normal/gamma mixture approximation to the full conditional posterior distribution, described in Subsections 5.4.3 through 5.4.5. For comparison reasons, we consider two approaches for the construction of the mixture approximation, regarding the derivation of the posterior moments: the proposed combined strategy involving both the entropy-based and the discretisation based techniques; and the latter method on its own. We will refer to the first approach as the 'combined' method.

For all the implementations of the Gibbs sampling algorithm in this application, we ran 3 independent chains, each of them having length $N = 2 \times 10^4$, resulting into a total of 6×10^4 iterations for each implementation. The main reason for that was to enable us to compute the Gelman and Rubin (1992) statistic, denoted by R, to assess convergence to the stationary distribution of the Markov chain. As mentioned earlier, the computation of the R statistic relies on

the use of multiple parallel sequences, started at points that are overdispersed with respect to the true posterior distribution. Then, the ratio of the estimated total variation of the posterior distribution of a parameter of interest, to the estimated within-sequence variation, both based on the second half of the simulated sequences, is examined. The R statistic is based on the square root of this ratio, and the authors show that it approaches 1 in the limit $N \to \infty$. Thus, values of R close to 1 suggest convergence to the distribution of interest. Here, the R statistic was calculated using CODA (Convergence Diagnosis and Output Analysis), a software package implementing S-PLUS functions for the analysis of MCMC output (e.g. Best, Cowles and Vines, 1995). CODA was also used to produce some of the diagnostic graphs appearing in this chapter.

To compare the results, we also analysed the data set using the Gibbs sampling method as implemented in BUGS (Spiegelhalter *et al.*, 1996). To tackle the problem of sampling from the nonstandard form of the full conditional posterior distribution of θ_i, the BUGS algorithm uses the adaptive rejection sampling technique developed by Gilks and Wild (1992), which can be applied since the density of interest is log-concave. The method is considered by the authors to be very efficient, however BUGS does not provide any information on the associated rejection rate. We stress here that our approach does not involve any rejection of simulated values, and is therefore more efficient than any method employing rejection sampling, at the expense of simulating from an approximate full conditional distribution. The case when an informative Inv-$\chi^2(\nu, \lambda)$ prior distribution for σ^2 is assumed, can be exactly specified and addressed with BUGS. However, when $\sigma^2 \sim U(0, \infty)$, BUGS requires a proper prior to be determined. Therefore, we assumed a Pareto$(1, 10^{-6})$ prior distribution for σ^{-2}, which is equivalent to a uniform $U(0, 10^6)$ prior specification on σ^2.

The posterior means and standard deviations of the parameters of interest were calculated as ergodic means of the MCMC output, whereas the 2.5% and 97.5% percentiles were obtained with appropriate ordering. As the Gelman and Rubin R statistic was very close to 1 after 3000 iterations of the algorithm for all the implementations of the method, the first 1500 simulated values of each chain were discarded. Figure 5.5 displays the monitored value of the Gelman and Rubin R diagnostic for the parameters θ_1, θ_4, θ_6, θ_8 and θ_9, when the combined method was used for the full conditional approximation, and model (4.1a) is assumed.

For all four implementations of the algorithm, convergence was also informally checked by visual inspection of the trace of the posterior mean of the parameters of interest, as the iterating process was developing. Figures 5.6 and 5.7 show the monitored posterior means for θ_1, θ_4, θ_6, θ_8 and θ_9, under model (4.1a), and

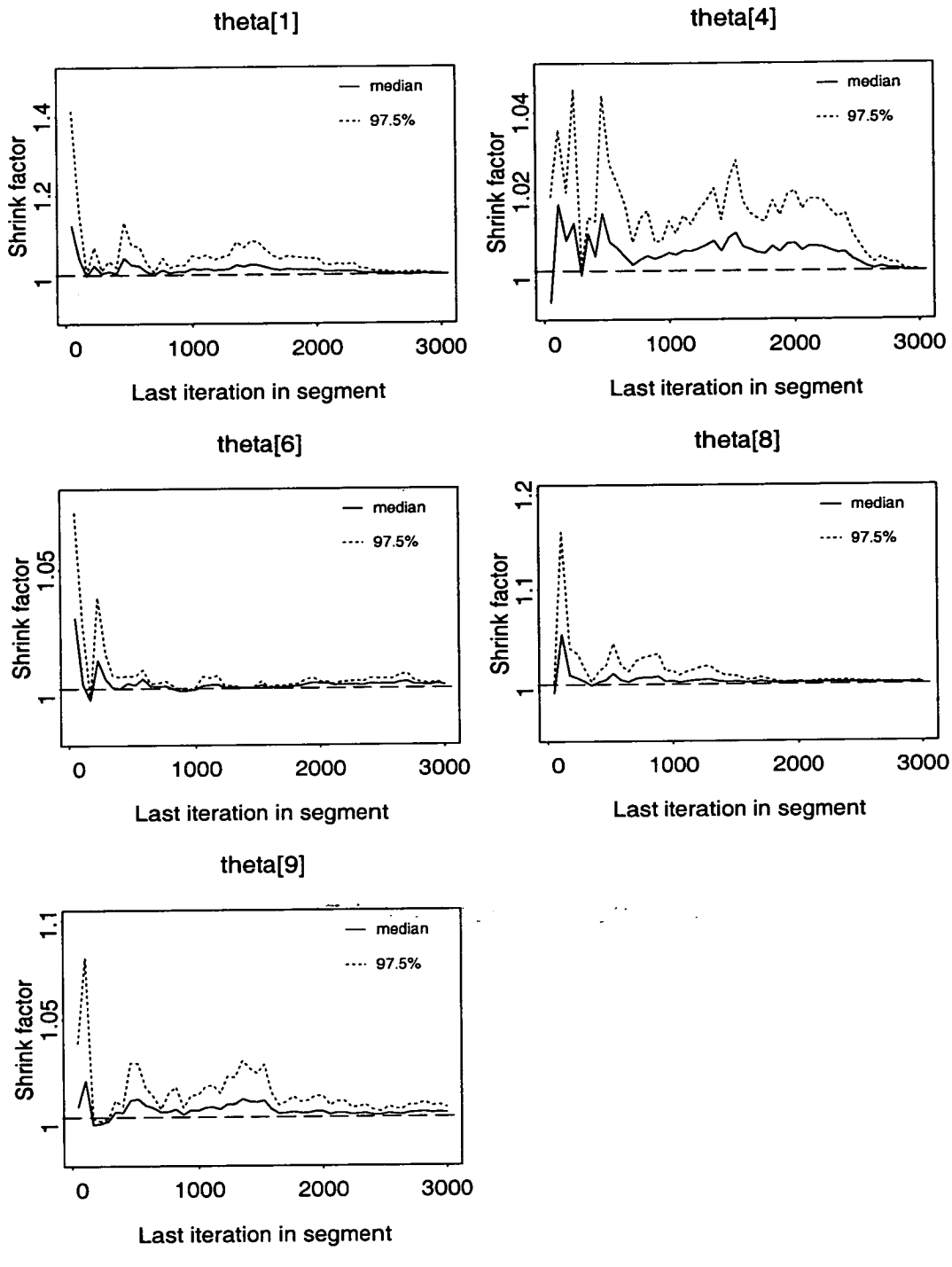

Figure 5.5: *Gelman and Rubin R statistic against iteration number for the parameters θ_1, θ_4, θ_6, θ_8 and θ_9 in the audit data example. Model (4.1a) is assumed and the entropy-discrete combined method is used for the approximation of* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$.

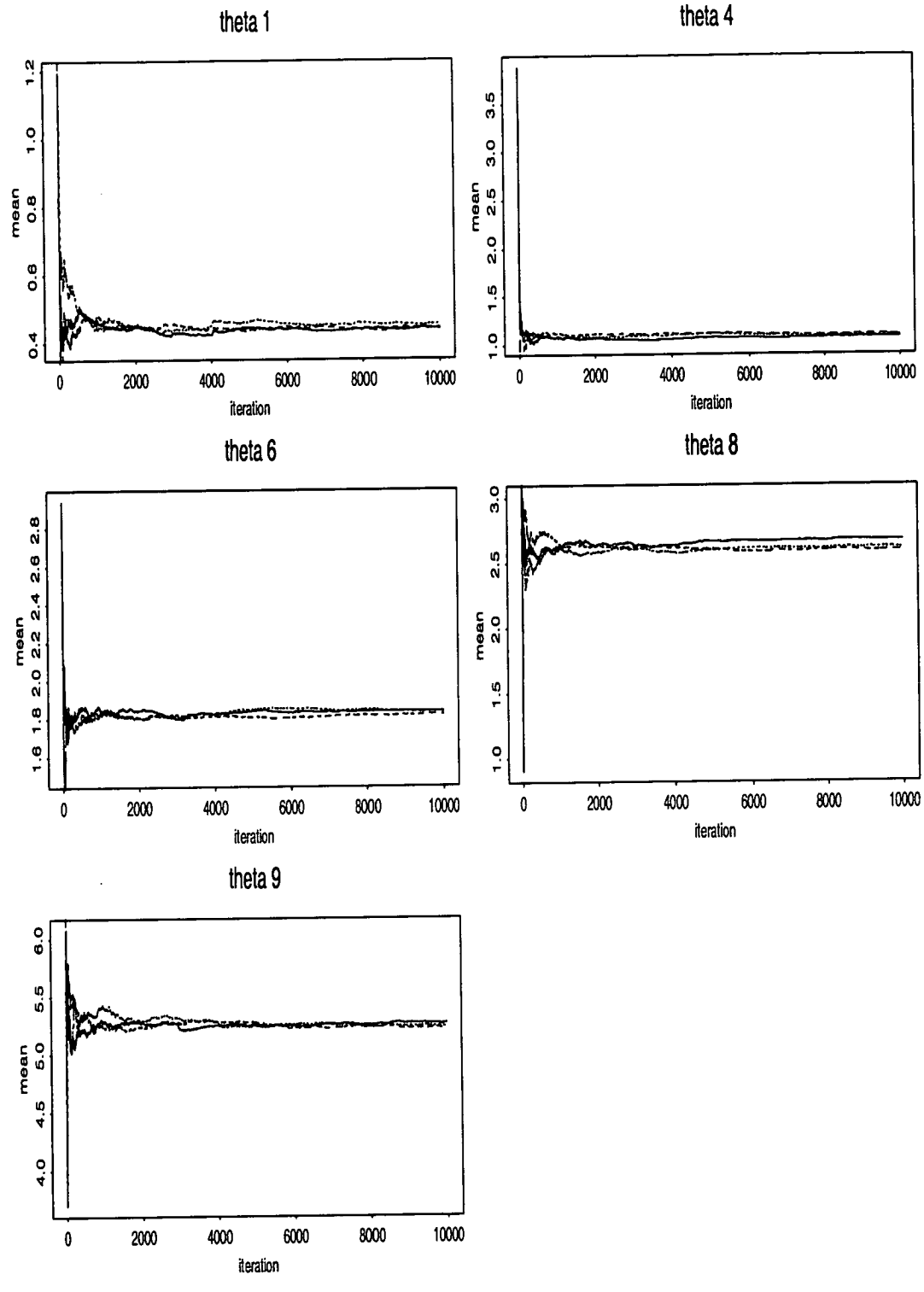

Figure 5.6: *Posterior mean trace for the parameters θ_1, θ_4, θ_6, θ_8 and θ_9 in the audit data example. Model (4.1a) is assumed and the entropy-discrete combined method is used for the approximation of $p(\theta_i|\mu, \sigma^2, \mathbf{y})$. The different lines correspond to the 3 independent Gibbs sampling chains.*

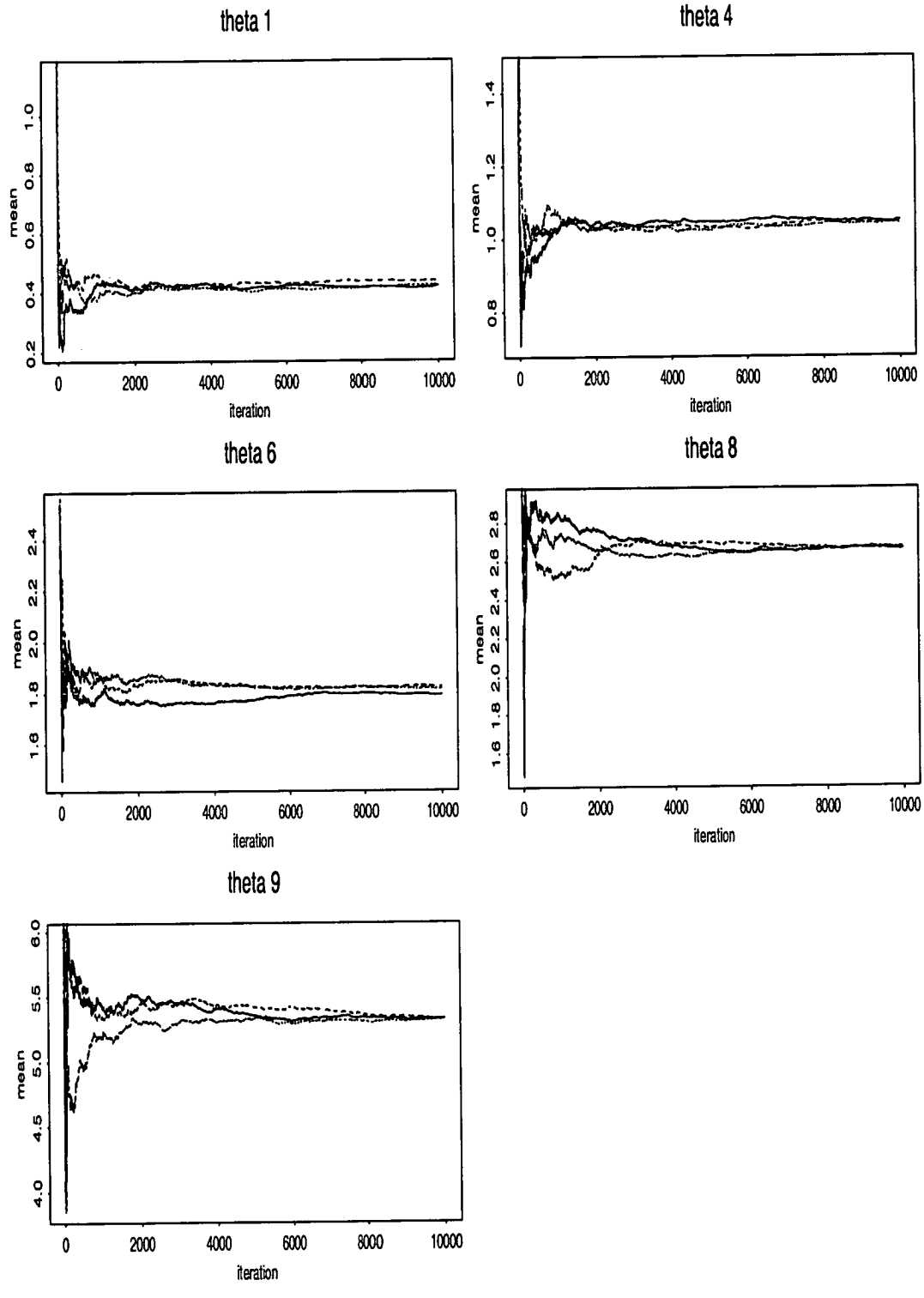

Figure 5.7: *Posterior mean trace for the parameters θ_1, θ_4, θ_6, θ_8 and θ_9 in the audit data example. Model (4.1a) is assumed and the discrete method is used for the approximation of $p(\theta_i|\mu, \sigma^2, \mathbf{y})$. The different lines correspond to the 3 independent Gibbs sampling chains.*

Table 5.5: *Approximate Gibbs estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for all audit data example parameters. Model (4.1a) is assumed, and the entropy-discrete combined and discrete-only methods are used for approximating* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$.

		Combined method				*Discrete method*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_3	0	0.46	0.58	0.00	2.01	0.43	0.56	0.00	1.98
θ_4, θ_5	1	1.09	0.93	0.06	3.48	1.04	0.96	0.02	3.46
θ_6, θ_7	2	1.84	1.25	0.27	5.01	1.81	1.44	0.10	5.37
θ_8	3	2.65	1.54	0.58	6.47	2.65	1.96	0.20	7.49
θ_9	6	5.26	2.32	1.75	10.70	5.28	3.68	0.63	14.50
μ		−0.37	1.06	−2.82	1.31	−0.52	1.15	−3.25	1.21
σ^2		6.84	11.20	0.29	31.50	7.41	11.00	0.30	34.20

for the combined and discretisation approximating methods respectively. The 3 different lines in each graph correspond to the 3 independent Gibbs sampling chains, started at different initial points.

The Gibbs sampling chains seem to converge to their approximate target distribution faster when the combined approximating method is preferred, especially when y is away from the origin, that is for θ_6, θ_8 and θ_9. However, similar plots showed that when a Inv-$\chi^2(\nu, \lambda)$ was assumed for the variance hyperparameter σ^2 under model (4.1b), there was no substantial difference between the two methods, due to some improvement to the discrete approximating technique.

Employing an approximation to enable simulation from the full conditional distribution of the Poisson means θ_i, raises the question of whether or not the Gibbs sampling chain converges to the correct target distribution. The results in Table 5.5 show that when model (4.1a) is assumed, the Gibbs sampler using the combined approximating technique produces estimates that are overall closer to the BUGS results in Table 5.6, than those obtained when the discretisation method is employed. In this case, the posterior mean estimates given from the latter method are relatively close to the BUGS estimates, although the estimation is rather poor as far as the standard deviation and the percentiles are concerned. On the other hand, when we use the combined approximation, all estimates, and especially those of the standard deviation and the percentiles are considerably

Table 5.6: *Gibbs sampling estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for all audit data example parameters using BUGS. Both model specifications (4.1a) and (4.1b) are considered.*

		Model (4.1a)				*Model (4.1b)*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_3	0	0.50	0.60	0.00	2.12	0.94	0.64	0.15	2.56
θ_4, θ_5	1	1.09	0.90	0.07	3.41	1.30	0.80	0.28	3.32
θ_6, θ_7	2	1.82	1.22	0.28	4.91	1.72	0.98	0.44	4.17
θ_8	3	2.59	1.52	0.59	6.37	2.20	1.17	0.63	5.07
θ_9	6	5.15	2.29	1.73	10.54	3.97	1.76	1.41	8.20
μ		−0.35	1.30	−3.16	1.21	0.22	0.39	−0.60	0.95
σ^2		8.04	29.31	0.26	44.73	0.59	0.30	0.24	1.36

Table 5.7: *Approximate Gibbs estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for all audit data example parameters. Model (4.1b) is assumed, and the entropy-discrete combined and discrete-only methods are used for approximating* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$.

		Combined method				*Discrete method*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_3	0	0.92	0.64	0.12	2.52	0.92	0.64	0.12	2.53
θ_4, θ_5	1	1.27	0.79	0.24	3.23	1.29	0.80	0.24	3.27
θ_6, θ_7	2	1.70	0.98	0.40	4.13	1.71	0.98	0.41	4.14
θ_8	3	2.18	1.18	0.58	5.11	2.21	1.19	0.60	5.13
θ_9	6	3.97	1.79	1.36	8.23	4.00	1.79	1.40	8.29
μ		0.19	0.42	−0.71	0.94	0.20	0.41	−0.68	0.95
σ^2		0.62	0.40	0.24	1.60	0.62	0.37	0.24	1.54

closer to these produced with BUGS. However, as Table 5.7 demonstrates, under the distributional assumptions given in model (4.1b), the two approximating methods perform equally well and they both produce estimates that are close to those given in Table 5.6. It is remarkable that both Gibbs sampling methods, and especially the one based on the discrete approximation, are considerably improved when more prior information on σ^2 is introduced in the model. This can be justified by the fact that assuming a more informative prior distribution for σ^2 has the effect of increasing the number of the degrees of freedom for the $\chi^2_{\nu+m}$ full conditional posterior distribution associated with σ^2, which in turn results in less extreme simulated values for this hyperparameter. Since our approximations to the full conditional posterior distribution of θ_i may be sensitive to extreme value combinations of μ, σ^2 and $\mathbf{y}$, we expect these approximating techniques to benefit from avoiding flat prior distributions.

5.4.7 Example: Oilwell discoveries data

In Subsections 5.4.4 and 5.4.5 it was emphasised that the approximations to the full conditional posterior distribution of θ_i are not very accurate when a zero observation is involved. However, as demonstrated with the audit data example in the preceding subsection, the Gibbs sampler using the $p(\theta_i|\mu, \sigma^2, \mathbf{y})$ approximating techniques can still perform well when zero counts are present in the data, especially when more prior knowledge is available for the hyperparameter σ^2. The performance of the method can be further improved when a large sample is available, as the increase in the sample size will again result in less variation in the σ^2 simulated values. We will use the oilwell discoveries data set, which was also considered in Section 3.7 and in Subsection 5.2.4, to illustrate this case. The data are reported in Table 3.3. As in Subsection 5.2.4, we assume that given the parameters $\theta_1, \theta_2, \ldots, \theta_{36}$, each of the observations $Y_1, Y_2, \ldots, Y_{36}$, independently follows a Poisson distribution with mean $\theta_1, \theta_2, \ldots, \theta_{36}$, respectively. At the first stage of the prior specification we assume that $\gamma_i = \log(\theta_i)$, $i = 1, \ldots, m$, are independently and identically distributed as normal $N(\mu, \sigma^2)$ random variables, whereas for the second stage, once more we consider two different settings: first we assume that μ and σ^2 are independently distributed according to vague uniform prior distributions over $(-\infty, \infty)$ and $(0, \infty)$ respectively; and then we assume the case where σ^2 has an Inv-$\chi^2(10, 0.46)$ distribution, independently from μ which again is uniformly distributed.

In this example we only report the results obtained employing the entropy-discrete combined technique for the approximation of the full conditional distribution of θ_i, which will be exclusively used hereafter. The simulation procedure was

Table 5.8: *Approximate Gibbs estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for the oilwell discoveries data example parameters. The entropy-discrete combined method is used for approximating* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$ *and both models (4.1a) and (4.1b) are considered.*

		Model (4.1a)				*Model (4.1b)*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_{19}	0	0.43	0.40	0.04	1.48	0.56	0.39	0.10	1.55
θ_{20}-θ_{29}	1	0.79	0.64	0.06	2.49	0.81	0.54	0.14	2.14
θ_{30}-θ_{33}	2	1.31	0.93	0.19	3.75	1.14	0.72	0.24	2.99
θ_{34}, θ_{35}	3	1.92	1.21	0.39	4.98	1.55	0.92	0.38	3.84
θ_{36}	5	3.33	1.76	0.86	7.49	2.54	1.33	0.75	5.86
μ		−0.76	0.32	−1.40	−0.15	−0.52	0.27	−1.09	−0.05
σ^2		1.29	0.65	0.25	2.80	0.58	0.27	0.25	1.26

the same as for the audit data example, except from the fact that the number of simulations for each of the implementations of the method was $N = 3 \times 10^4$, with 3 independent chains of length 10^4 for each implementation. This was dictated by the large size of the data set and computer memory and storage reasons.

The results are presented in Table 5.8 for both model specifications (4.1a) and (4.1b). As illustrated in Figure 5.8 for model (4.1a), the Gelman and Rubin R statistic was virtually equal to 1 after 1000 iterations for both implementations of the method. The convergence of the algorithm was also indicated by monitored ergodic means of the Gibbs sampling output. Figures 5.9 and 5.10 display the trace of the posterior means for five θ parameters corresponding to distinct y values, namely θ_1, θ_{20}, θ_{30}, θ_{34} and θ_{36}, under models (4.1a) and (4.1b) respectively. The monitored means from the 3 independent chains seem to be very close after about 3000 simulations for model (4.1a) and 1000 simulations for model (4.1b), suggesting that convergence seems to occur faster when the scaled inverse chi-square prior distribution is used under model (4.1b). The burn-in period was specified accordingly for the two cases.

For comparison reasons, we also report the estimates obtained with BUGS in Table 5.9. The results show that under model (4.1a), there is some disagreement between the estimates produced by the approximate Gibbs method and BUGS. However, given the high number of zero values in the data, we notice that the

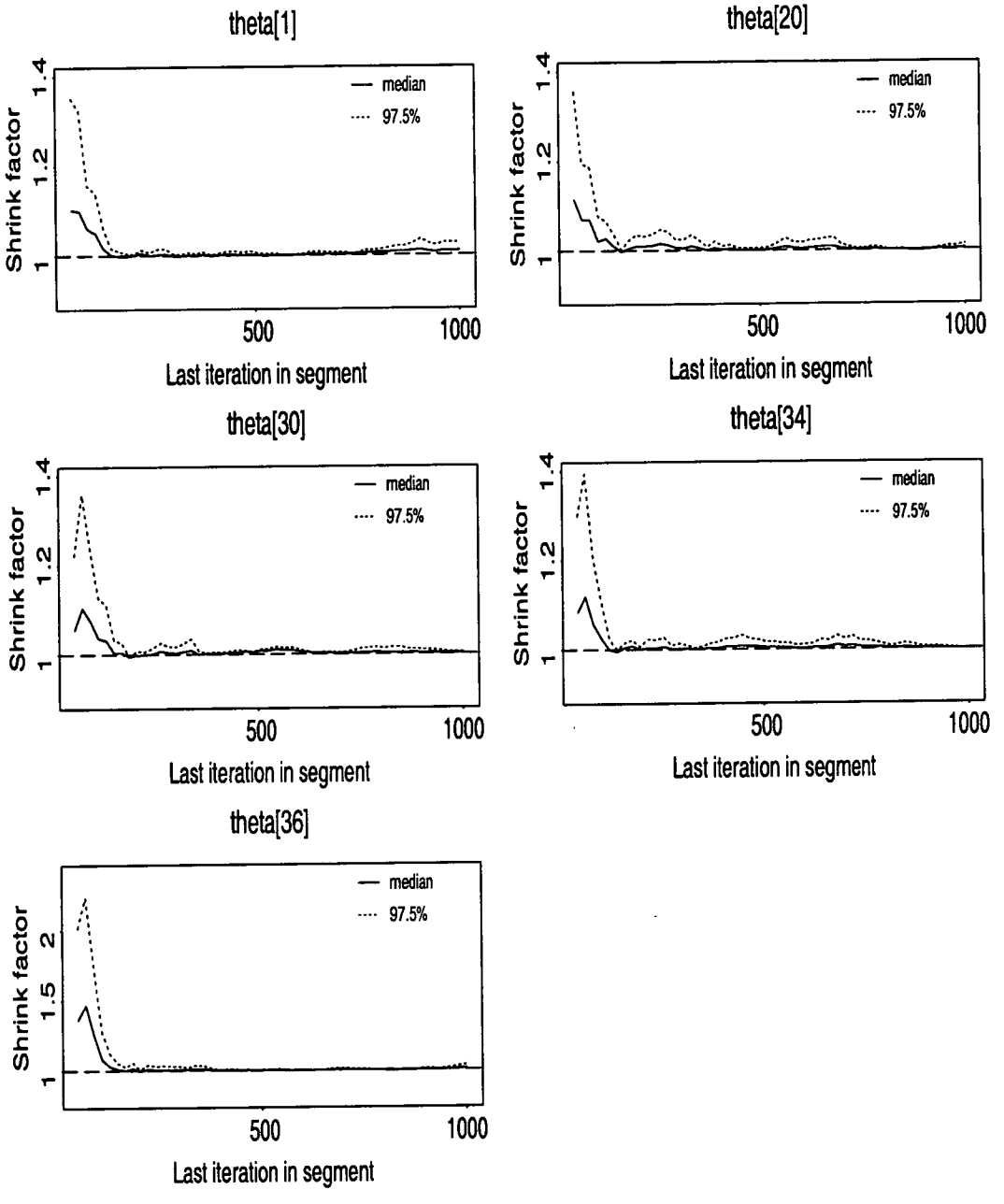

Figure 5.8: *Gelman and Rubin R statistic against iteration number for the parameters θ_1, θ_{20}, θ_{30}, θ_{34} and θ_{36} in the oilwell discoveries data example. Model (4.1a) is assumed and the entropy-discrete combined method is used for the approximation of $p(\theta_i|\mu, \sigma^2, \mathbf{y})$.*

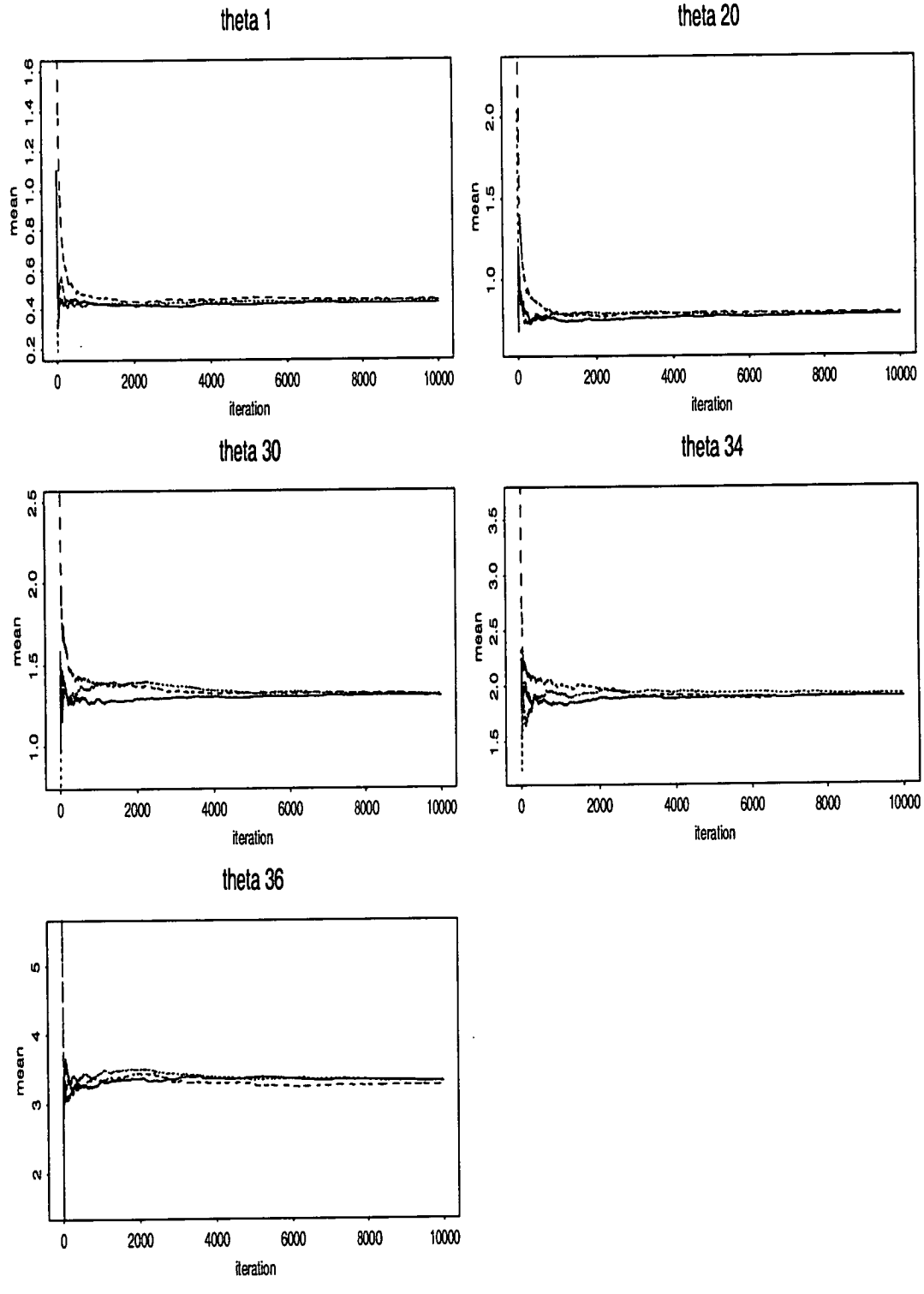

Figure 5.9: *Posterior mean trace for the parameters* θ_1, θ_{20}, θ_{30}, θ_{34} *and* θ_{36} *in the oilwell discoveries data example. Model (4.1a) is assumed and the entropy-discrete combined method is used for the approximation of* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$. *The different lines correspond to the* 3 *independent Gibbs sampling chains.*

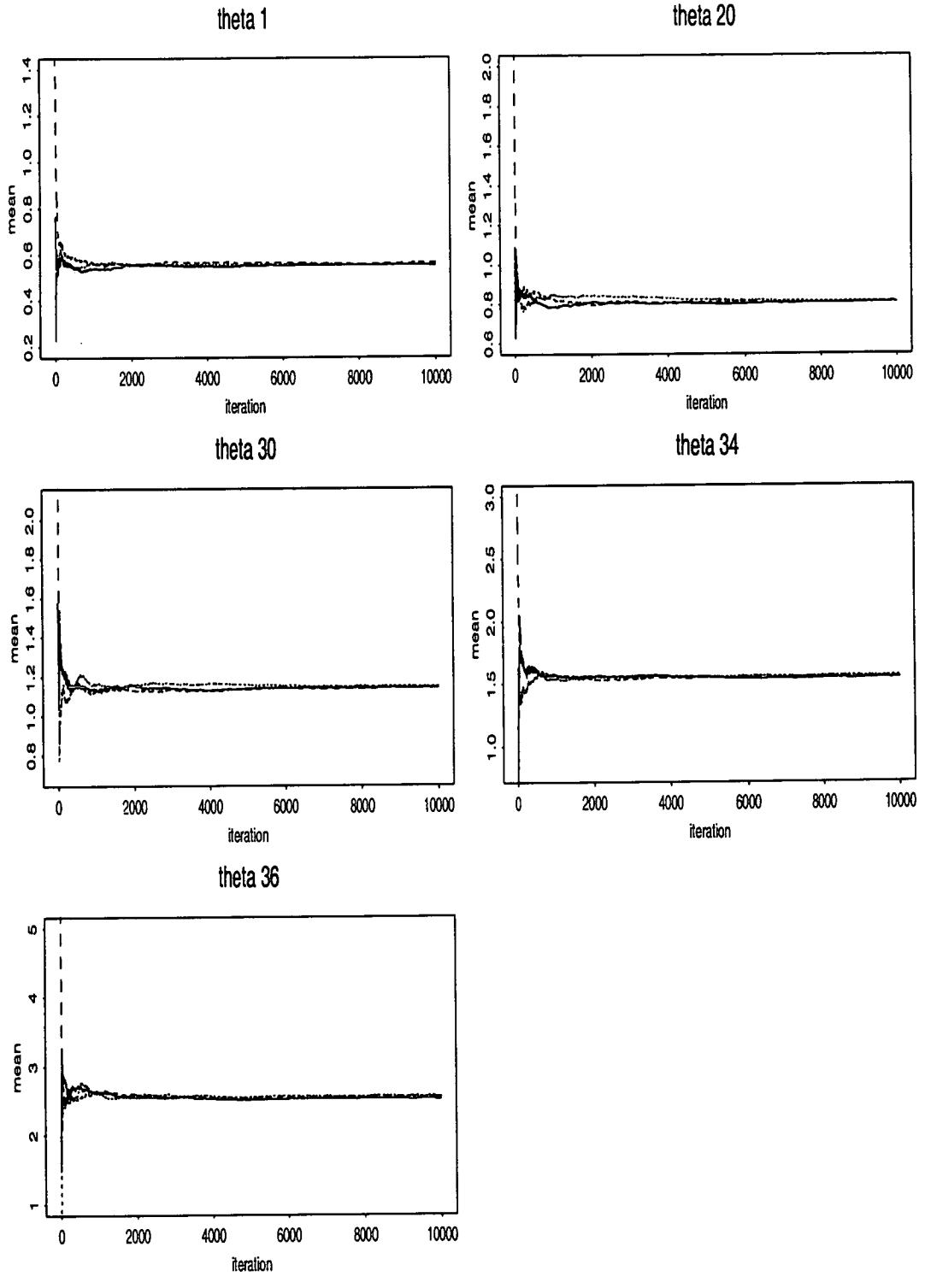

Figure 5.10: *Posterior mean trace for the parameters* θ_1, θ_{20}, θ_{30}, θ_{34} *and* θ_{36} *in the oilwell discoveries data example. Model (4.1b) is assumed and the entropy-discrete combined method is used for the approximation of* $p(\theta_i|\mu, \sigma^2, \mathbf{y})$. *The different lines correspond to the* 3 *independent Gibbs sampling chains.*

Table 5.9: *Gibbs sampling estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for the oilwell discoveries data example parameters using BUGS. Both models (4.1a) and (4.1b) are considered.*

		Model (4.1a)				*Model (4.1b)*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_{19}	0	0.46	0.41	0.03	1.54	0.56	0.39	0.09	1.58
θ_{20}-θ_{29}	1	0.82	0.64	0.11	2.49	0.82	0.54	0.17	2.18
θ_{30}-θ_{33}	2	1.29	0.90	0.24	3.64	1.15	0.71	0.28	2.98
θ_{34}, θ_{35}	3	1.85	1.18	0.43	4.90	1.54	0.90	0.42	3.82
θ_{36}	5	3.16	1.74	0.83	7.45	2.52	1.32	0.76	5.82
μ		−0.71	0.37	−1.55	−0.09	−0.51	0.26	−1.05	−0.04
σ^2		1.25	0.96	0.16	3.71	0.57	0.26	0.24	1.22

approximate method performs reasonably well, and the accuracy of the estimates is comparable with that of the audit data estimates for the same model assumptions. This is due to the effect of the larger size of the oilwell data set. The effect is even more apparent under model (4.1b), in which case the estimates obtained with the approximate method are very close to the BUGS results, as shown in Tables 5.8 and 5.9.

5.5 A hybrid MCMC method

The analysis of the examples in the preceding section suggests that despite the fact that the approximations to the full conditional distribution usually perform well, they can also exhibit low accuracy in some cases, especially with data sets of small size, or containing a large number of zero observations. As a result, the Gibbs sampler chain may converge to a distribution other than the one under consideration, and therefore produce wrong estimates.

Hence, we need to adopt an alternative method for the Bayesian analysis of our hierarchical model. The problem of sampling from full conditional distributions that are not given in standard analytical form, may be addressed in various ways, including those mentioned in Subsection 5.3.5. However, some of these methods can be considerably inefficient to implement in terms of either setting-up effort, or computing time when embedded within a Gibbs sampling algorithm. One al-

ternative strategy is combining the Gibbs sampler with other MCMC techniques, and thus adopting a so-called hybrid MCMC approach.

5.5.1 The Metropolis-Hastings within Gibbs method

The Metropolis-Hastings algorithm, described in Subsection 5.3.3, requires the knowledge of the target density only up to the constant of normalisation. It then proceeds with sampling a candidate point from a proposal distribution, which is accepted with a probability whose evaluation involves the non-normalised target and the proposal density, as shown in Algorithm 5.1 of Subsection 5.3.3. This suggests that when one or more of the full conditional distributions needed for the Gibbs sampler are only available in a non-normalised form, one may employ a Metropolis-Hastings subalgorithm within a Gibbs sampler step to obtain the necessary simulated value at the current iteration. The target density for the Metropolis-Hastings algorithm will be the nonclosed form of the full conditional posterior distribution and, in theory, any proposal distribution may be used to generate the candidate variates. The Metropolis-Hastings subalgorithm can itself comprise T steps. Despite the fact that neither of the two combined chains would correctly converge to the right distribution if individually applied, the convergence issue is not affected by the combination of the two MCMC techniques. The hybrid method will still converge to the target distribution, given that the combined chain is irreducible and aperiodic (Tierney, 1994).

The remaining questions are how to choose the number T of the Metropolis-Hastings substeps, and the proposal distribution for the Metropolis-Hastings algorithm. The answer to the former is that one substep suffices, as convergence will occur for any value of T. Thus, the choice depends on the nature of the specific problem, although taking $T = 1$ has become common practice. Selecting the proposal distribution also depends on the problem under consideration, and usually the statistician will have to compromise between efficiency of evaluation and acceleration of mixing for the Metropolis-Hastings algorithm.

The idea first appears in Muller (1991), and has since become widely known as the Metropolis-Hastings within Gibbs method. Some authors (e.g., Chib and Greenberg, 1995) disagree with this term, since the whole method can be viewed as a single-component Metropolis-Hastings algorithm, described in Section 5.3.4, using as its proposal distributions the full conditional distributions when the latter are fully available, and employing some other candidate density when the full conditional distribution is known up to a normalising constant. In that sense, the term 'Gibbs within Metropolis-Hastings', or simply 'single-component Metropolis-Hastings', used by some authors, would seem more appropriate.

5.5.2 The Metropolis-Hastings within Gibbs method for the Poisson/log-normal model

For the Poisson/log-normal models (4.1a) and (4.1b), the full conditional posterior distribution of θ_i, $i = 1, \ldots, m$, is given in nonclosed analytical form, as shown in (5.48). Adopting the Metropolis-Hastings within Gibbs approach, and following the methodology described in the preceding section we can proceed as follows. At iteration t we employ two Gibbs steps to sample from the full conditional posterior distributions of μ and σ^2, which involves straightforward simulation from a normal and a chi-square distribution respectively. Then, for each $i = 1, \ldots, m$, we generate a candidate θ_i from a chosen proposal distribution and we accept the new point with a probability given by (5.35). If we denote the proposal distribution by $q(\cdot)$ and the full conditional posterior distributions by $p(\cdot \mid \cdot)$, for which we will suppress dependence on $\mathbf{y}$, we can now give an outline of the Metropolis-Hastings within Gibbs algorithm for the Poisson/log-normal model.

ALGORITHM 5.3: *METROPOLIS-HASTINGS WITHIN GIBBS.*

1. *Choose starting values* $\left(\mu^{(0)}, \sigma^{2(0)}, \theta_1^{(0)}, \ldots, \theta_m^{(0)}\right)$
2. *For* $t = 1, 2, 3, \ldots$ *(until convergence):*
   - *(a) At iteration t, take as input the point* $\left(\mu^{(t-1)}, \sigma^{2(t-1)}, \theta_1^{(t-1)}, \ldots, \theta_m^{(t-1)}\right)$
   - *(b) Generate* $\mu^{(t)}$ *from* $N\left(\frac{\sum_{i=1}^m log\,\theta_i^{(t-1)}}{m}, \frac{\sigma^{2(t-1)}}{m}\right)$
   - *(c) Generate* $\sigma^{2(t)}$ *such that* $\frac{\sum_{i=1}^m \left\{log\,\theta_i^{(t-1)} - \mu^{(t)}\right\}^2}{\sigma^{2(t)}} \sim \chi^2_{m-2}$ *or* $\frac{\nu\lambda + \sum_{i=1}^m \left\{log\,\theta_i^{(t-1)} - \mu^{(t)}\right\}^2}{\sigma^{2(t)}} \sim \chi^2_{\nu+m}$
   - *(d) For* $i = 1$ *to* m*:*
     - *Generate* $\theta_i^{(t)}$ *from* $q_i\left(\cdot \mid \mu^{(t)}, \sigma^{2(t)}, \boldsymbol{\theta}^{(t-1)}\right)$
     - *Calculate the ratio* $r = \dfrac{p\left(\theta_i^{(t)} \middle| \mu^{(t)}, \sigma^{2(t)}\right) q_i\left(\theta_i^{(t-1)} \middle| \mu^{(t)}, \sigma^{2(t)}, \theta_i^{(t)}, \boldsymbol{\theta}_{-i}^{(t-1)}\right)}{p\left(\theta_i^{(t-1)} \middle| \mu^{(t)}, \sigma^{2(t)}\right) q_i\left(\theta_i^{(t)} \middle| \mu^{(t)}, \sigma^{2(t)}, \theta_i^{(t-1)}, \boldsymbol{\theta}_{-i}^{(t-1)}\right)}$
     - *Set* $\theta_i^{(t)} = \begin{cases} \theta_i^{(t)} & \text{with probability } min(1, r) \\ \theta_i^{(t-1)} & \text{otherwise} \end{cases}$

Here, $\boldsymbol{\theta}_{-i}^{(t-1)}$ represents the vector of all components of $\boldsymbol{\theta}$, excluding θ_i, at their

current values, that is

$$\boldsymbol{\theta}_{-i}^{(t-1)} = \left(\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)}, \ldots, \theta_m^{(t-1)}\right).$$

We also notice that the full conditional posterior distribution of θ_i does not depend on the remaining parameters of vector $\boldsymbol{\theta}$, and thus, $\boldsymbol{\theta}_{-i}^{(t-1)}$ does not appear in the notation for $p(\cdot \mid \cdot)$.

We now need to choose the proposal distribution for the Metropolis-Hastings subalgorithm. Tierney (1994) suggests an independence type Metropolis-Hastings scheme, that is he discusses the use of independent proposal distributions. Here, by independence we mean that the proposal distribution of θ_i at iteration t does not depend on the iteration $(t-1)$ value of θ_i, and vice versa. Then, the acceptance ratio in step $2(d)$ of Algorithm 5.3 becomes

$$r = \frac{w\left(\theta_i^{(t)}\right)}{w\left(\theta_i^{(t-1)}\right)}, \tag{5.84}$$

where $w\left(\theta_i^{(t)}\right)$ is the ratio of the full conditional posterior density of $\theta_i^{(t)}$ and the proposal density at $\theta_i^{(t)}$, that is

$$w\left(\theta_i^{(t)}\right) = \frac{p\left(\theta_i^{(t)} \middle| \mu^{(t)}, \sigma^{2(t)}\right)}{q_i\left(\theta_i^{(t)} \middle| \mu^{(t)}, \sigma^{2(t)}, \boldsymbol{\theta}_{-i}^{(t-1)}\right)}. \tag{5.85}$$

Hence, we notice that the acceptance probability of the Metropolis-Hastings subalgorithm is given as a ratio of importance weights, where $p(\cdot)$ can be viewed as the target density and $q(\cdot)$ as the importance function. In that sense, following the principles of the importance sampling technique, it seems reasonable to choose a proposal density $q(\cdot)$ which is a good approximation to the full conditional density $p(\cdot)$. The more similar that the two densities are, the closer the ratio r is to 1, and consequently the generated θ_i values are more likely to be accepted.

Following this line of thought, we consider two possible proposal distributions, that both approximate the full conditional posterior distribution of θ_i: a $\text{Ga}(y_i + 1, 1)$ distribution, and the log-normal/gamma mixture distribution developed in Subsections 5.4.3 through 5.4.5. The former was employed as the importance density, approximating $p(\theta_i|\mu, \sigma^2, \mathbf{y})$, in the implementation of the importance sampling technique to the estimation of the conditional expectation $\text{E}(\theta_i|\mu, \sigma^2, \mathbf{y})$ in Chapter 3. The mixture log-normal/gamma distribution was shown to provide an even better approximation to the full conditional posterior distribution of the Poisson means. However, we notice here that using a good approximation to the full conditional distribution as our proposal distribution, will only affect the

efficiency and mixing speed of the Metropolis-Hastings within Gibbs algorithm, since any other choice would also eventually lead to the convergence of the hybrid chain to its stationary distribution.

5.5.3 Example: Audit data

In the present subsection we illustrate the performance of the Metropolis-Hastings within Gibbs method, reanalysing the audit data set also considered before. In Subsection 5.4.6 we estimated the Poisson means in the audit data example, employing a Gibbs sampling scheme which was based on approximate conditional posterior distributions for the parameters θ_i, $i = 1, \ldots, m$. Now, the Metropolis-Hastings within Gibbs method, presented in Subsection 5.5.2, enables us to obtain the exact full hierarchical analysis of the data. Once more, we employ the same model specification as in Subsections 5.2.3 and 5.4.6.

We follow Algorithm 5.3. As proposed in Subsection 5.5.2, two different proposal distributions will be used in step $2(d)$ of the algorithm to provide a candidate point for the Metropolis-Hastings subalgorithm: the log-normal/gamma mixture distribution, featuring the enropy-discrete combined moment approximating method, presented in Subsections 5.4.3 through 5.4.5, and a gamma distribution with shape parameter equal to $(y_i + 1)$ and scale equal to 1. We will refer to the former as the mixture proposal and to the latter as the gamma proposal. As with the approximate Gibbs method, 3 independent chains, each of size $N = 2 \times 10^4$, will permit the evaluation of the Gelman and Rubin R statistic for the assessment of the convergence of the algorithm. However, this time the burn-in period will vary according to the selected prior setting and the proposal distribution. That reflects the effect that both the model specification and the choice of the proposal distribution have on the mixing speed of the method. We will consider the two specifications for the variance parameter σ^2 under models (4.1a) and (4.1b) separately.

Model (4.1a): Uniform hyperprior on σ^2

In Table 5.10 we report the estimates for the posterior mean, standard deviation and 2.5% and 97.5% percentiles of all the model parameters, for the case when model (4.1a) is assumed, and for both the mixture and gamma proposal distribution choices. When the mixture proposal was employed, the acceptance rate for the Metropolis-Hastings subalgorithm was between 75% and 99%, with the lowest rate being for the θ parameters corresponding to the zero data values. This is an exceptionally high rate for a Metropolis-Hastings application, and apparently reflects the high accuracy of the mixture approximation to the full conditional

Table 5.10: *Metropolis-Hastings within Gibbs estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for all audit data example parameters. Model (4.1a) is assumed, and both mixture and gamma proposals are considered for the Metropolis-Hastings subalgorithm.*

		Mixture proposal				*Gamma proposal*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_3	0	0.50	0.60	0.00	2.13	0.49	0.58	0.00	2.05
θ_4, θ_5	1	1.09	0.91	0.07	3.41	1.09	0.90	0.08	3.40
θ_6, θ_7	2	1.81	1.22	0.29	4.91	1.80	1.23	0.29	4.95
θ_8	3	2.59	1.52	0.59	6.36	2.58	1.53	0.57	6.40
θ_9	6	5.14	2.29	1.71	10.50	5.15	2.30	1.55	10.50
μ		−0.29	1.01	−2.72	1.19	−0.31	1.00	−2.75	1.20
σ^2		5.93	12.90	0.24	31.10	6.01	11.50	0.24	31.10

posterior distribution of interest, especially when y is not close to the origin. The Gelman and Rubin R statistic after 8×10^3 iterations of the algorithm, was not higher than 1.05 for any of the model parameters. This suggested that the first 4000 iterations should be discarded, although the choice of the burn-in period was complicated by the high variation exhibited in the σ^2 simulations. This is shown in Figure 5.11, which displays the trace of the simulated values of the model parameters for the first 8000 iterations of all 3 independent chains. We attribute this remarkably high variation to the uniform $U(0, \infty)$ prior distribution that we assumed for the variance hyperparameter σ^2. As shown before, this prior setting results in a χ^2_{m-2} full conditional distribution for $\frac{\sum_{i=1}^{m}(\gamma_i-\mu)^2}{\sigma^2}$, and thus, occasional draws that are very close to zero give very large values for σ^2.

We notice that the estimates for the Poisson means θ_i in the left half of Table 5.10 are almost identical to those obtained with BUGS in Table 5.6. This suggests that the Metropolis-Hastings within Gibbs method gives the exact results, correcting the inaccuracy occurred when an approximation to the full conditional posterior distribution of θ_i was used with the Gibbs sampling method. However, there is some discrepancy in the estimates of the hyperparameters μ and σ^2. We believe that this is partly due to the effect of the high variation in the σ^2 sampling. Nevertheless, our experimentation with assuming different ranges for the uniform σ^2 distribution involved in the BUGS model specification, indicated that

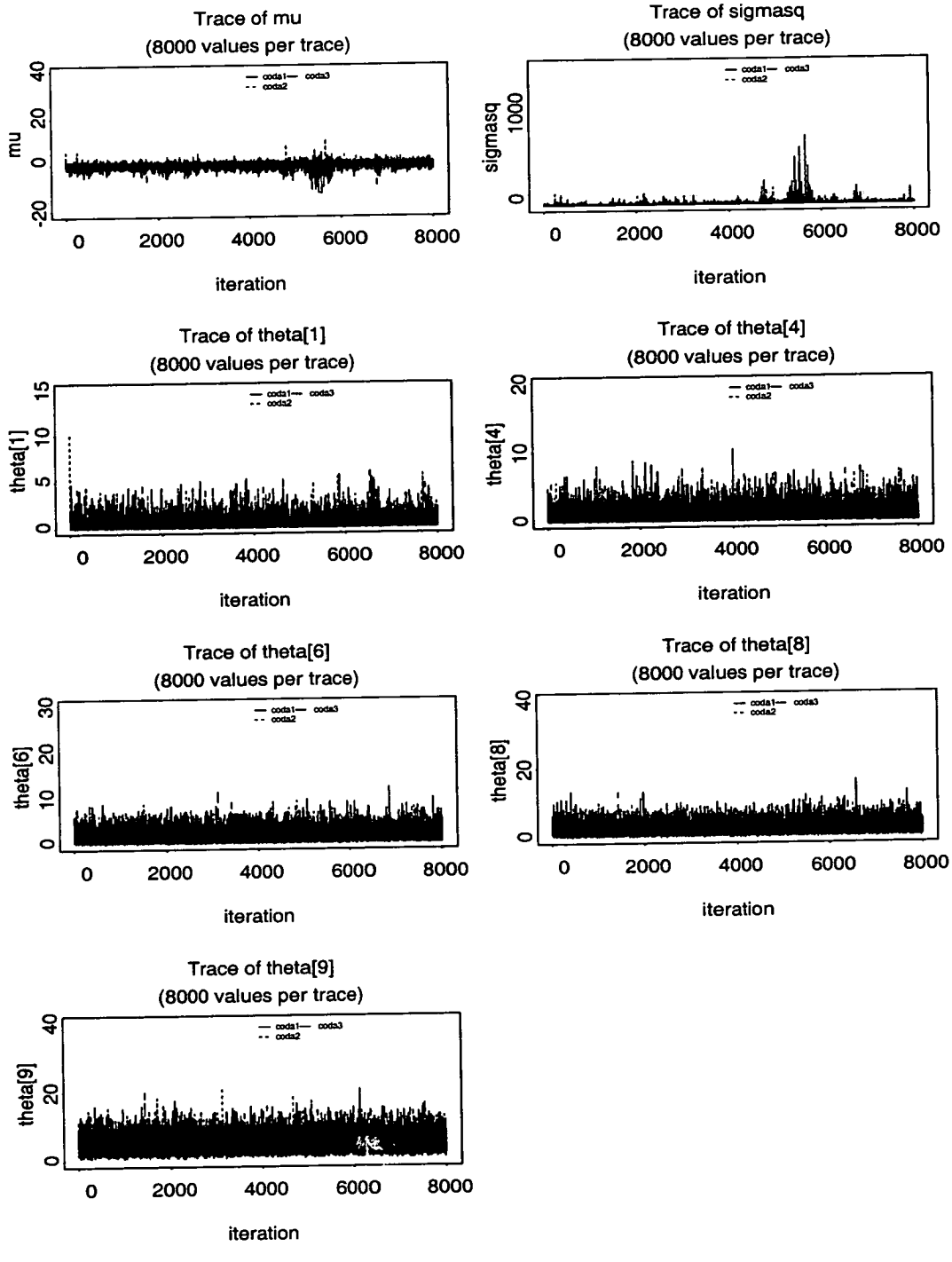

Figure 5.11: *Trace of sampled values from 3 parallel chains for the audit data example parameters. Model (4.1a) is assumed and a mixture proposal is used for the Metropolis-Hastings subalgorithm.*

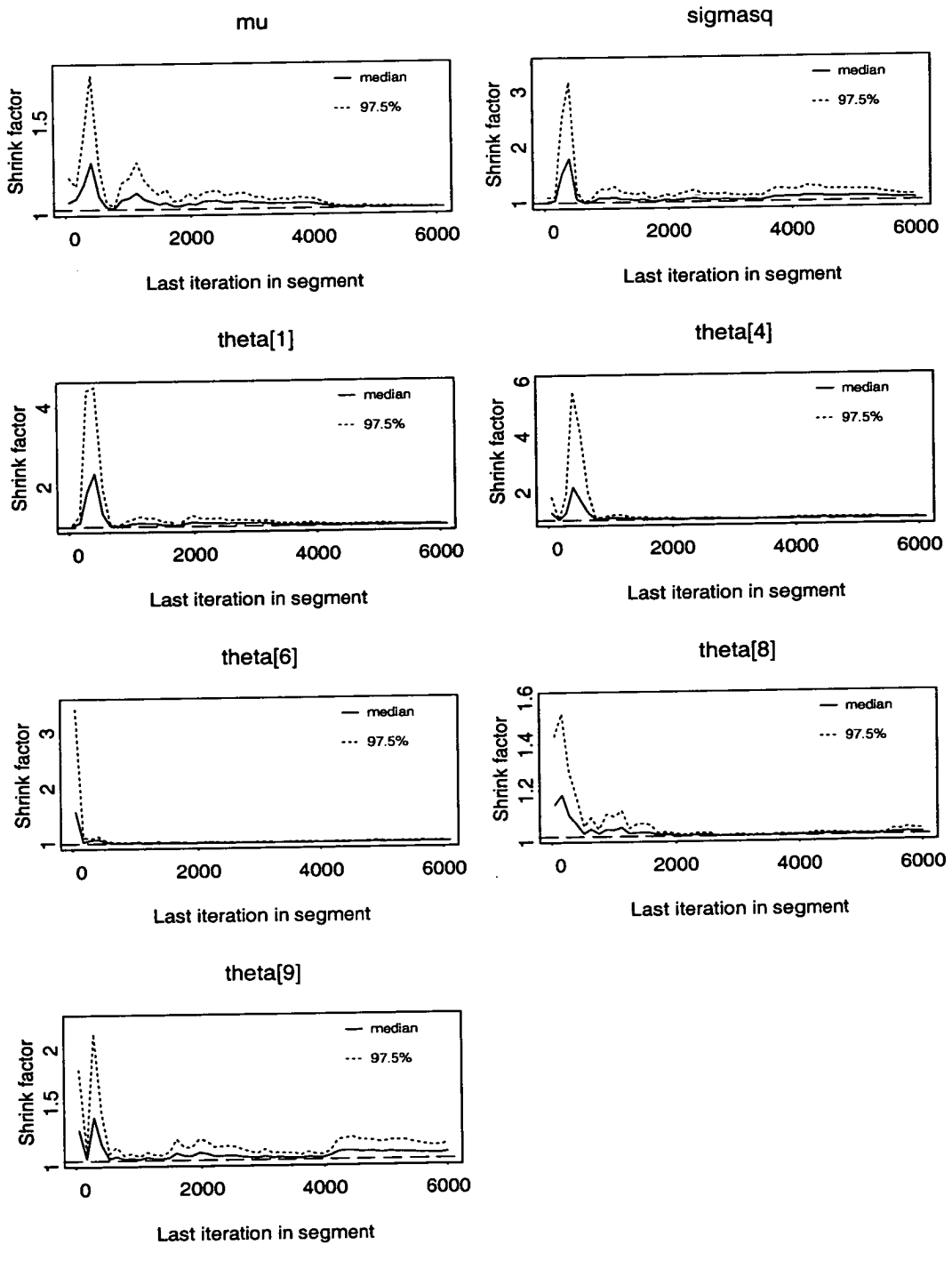

Figure 5.12: *Gelman and Rubin R statistic against iteration number for the audit data example parameters. Model (4.1a) is assumed and the gamma proposal is used for the Metropolis-Hastings subalgorithm.*

this discrepancy could also be caused by the fact that the BUGS estimates were quite sensitive to the selection of the σ^2 uniform prior distribution.

We then employed a $\text{Ga}(y_i+1, 1)$ proposal density for the Metropolis-Hastings subalgorithm. The acceptance rate dropped to 49%-59%. This was expected, since the gamma proposal does not approximate the full conditional distribution of θ_i as accurately as the log-normal/gamma mixture distribution does. However, this acceptance rate can still be considered very satisfactory. Relying on the Gelman and Rubin R diagnostic, we ignored the first 3000 simulations of each chain, since the R statistic was practically equal to 1 after 6000 iterations of the algorithm. This is illustrated in Figure 5.12, where the value of the R statistic is monitored during the first 6000 iterations. Also, due to the flat uniform prior used on σ^2, the simulated values for this hyperparameter again exhibit high variation. The estimates reported on the right hand side of Table 5.10 are very close to those obtained when the mixture proposal was used. However, in a few cases (e.g. for the percentile estimates), the estimates obtained using the mixture proposal seem to be slightly more consistent with the BUGS results, implying that maybe longer chains should be needed with the gamma proposal, to overcome the effect of a lower acceptance rate.

Model (4.1b): Inv-$\chi^2(\nu, \lambda)$ hyperprior on σ^2

We now assume that the variance parameter σ^2 follows a Inv-$\chi^2(\nu, \lambda)$ distribution, according to the model specification (4.1b). We take the parameters of this distribution to be $\nu = 10$ and $\lambda = 0.45$, as we did in Subsection 5.4.6. When the mixture proposal is preferred, an outstandingly high acceptance rate of 98%-99% suggests that the mixture proposal density is now almost identical to the full conditional distribution under consideration. The Gelman and Rubin R statistic did not take more than 3000 iterations to settle down to 1, as again demonstrated in Figure 5.13. We therefore discarded the first 1500 iterations from the output of each chain. Furthermore, comparing Figure 5.14 to Figure 5.11, we notice that the variation of the σ^2 simulated values has now dropped to normal levels, as a result of the introduced prior information for this parameter. The estimates of the model parameters using both proposal distributions, are presented in Table 5.11. The results reported on the left hand side of the table show that, in this case, our Metropolis-Hastings within Gibbs method and BUGS essentially produce the same estimates for all the model parameters, including the hyperparameters μ and σ^2, for which there was some disagreement when a uniform prior was assumed for σ^2.

Finally, we also ran the Metropolis-Hastings within Gibbs algorithm with the $\text{Ga}(y_i + 1, 1)$ candidate distribution, for the case when an Inv-$\chi^2(10, 0.45)$

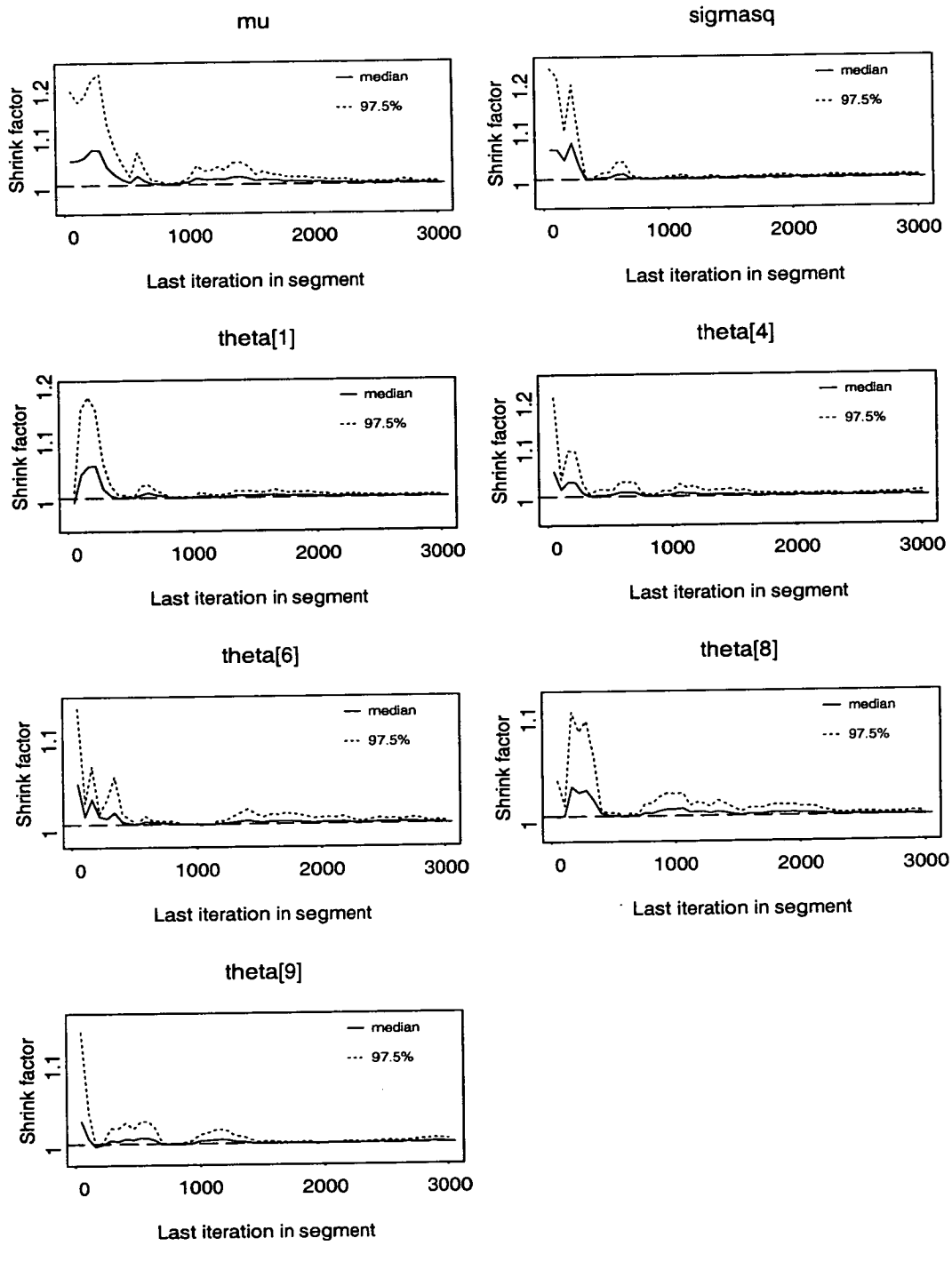

Figure 5.13: *Gelman and Rubin R statistic against iteration number for the audit data example parameters. Model (4.1b) is assumed and the mixture proposal is used for the Metropolis-Hastings subalgorithm.*

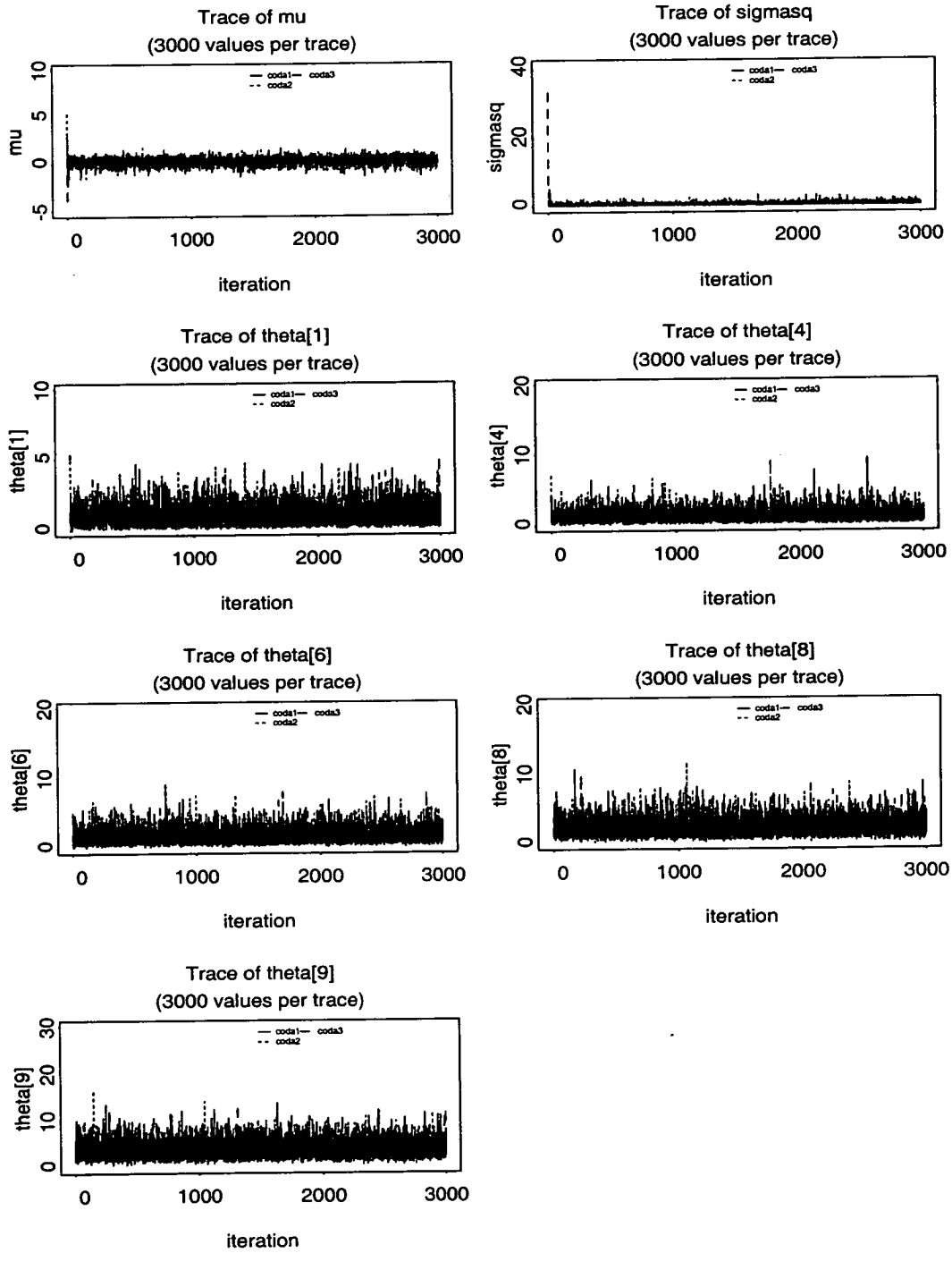

Figure 5.14: *Trace of sampled values from 3 parallel chains for the audit data example parameters. Model (4.1b) is assumed and the mixture proposal is used for the Metropolis-Hastings subalgorithm.*

Table 5.11: *Metropolis-Hastings within Gibbs estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for the audit data example parameters. Model (4.1b) is assumed and both mixture and gamma proposals are considered for the Metropolis-Hastings subalgorithm.*

		Mixture proposal				*Gamma proposal*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_3	0	0.94	0.64	0.16	2.56	0.94	0.63	0.16	2.55
θ_4, θ_5	1	1.29	0.80	0.28	3.32	1.29	0.79	0.27	3.28
θ_6, θ_7	2	1.72	0.98	0.44	4.18	1.71	0.97	0.44	4.14
θ_8	3	2.20	1.17	0.63	5.10	2.18	1.17	0.63	5.08
θ_9	6	3.97	1.77	1.41	8.19	3.97	1.76	1.42	8.13
μ		0.22	0.40	−0.60	0.94	0.22	0.39	−0.61	0.94
σ^2		0.58	0.30	0.23	1.34	0.58	0.30	0.24	1.34

prior distribution is assumed for σ^2. The estimates, given in Table 5.11, are almost identical to those obtained with the mixture proposal, with only a few slight discrepancies in the second decimal place. Again, the Gelman and Rubin diagnostic was below 1.05 after 3000 iterations. However, as the acceptance rate decreases to 31%-60%, we expect that a larger number of simulations might be required to achieve a certain degree of accuracy.

5.5.4 Example: Oilwell discoveries data

We followed the same procedure to analyse the oilwell discoveries data using the Metropolis-Hastings within Gibbs algorithm. The data set is given in Table 3.3. We use the same model specification as in Subsection 5.4.7.

Model (4.1a): Uniform hyperprior on σ^2

In Subsection 5.4.7, it was demonstrated that when the mixture approximation to $p(\theta_i|\mu, \sigma^2, \mathbf{y})$ is employed in its own right as the actual full conditional posterior distribution of θ_i, the Gibbs sampler chain may converge to a stationary distribution which is not exactly the conditional distribution under consideration. However, when this approximation is embedded in the Metropolis-Hastings within Gibbs method, used as the proposal distribution for the Metropolis-Hastings sub-

Table 5.12: *Metropolis-Hastings within Gibbs estimates of the posterior mean, standard deviation and (2.5%, 97.5%) percentiles for all oilwell discoveries data example parameters. Model (4.1a) is assumed, and both mixture and gamma proposals are considered for the Metropolis-Hastings subalgorithm.*

		Mixture proposal				*Gamma proposal*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_{19}	0	0.46	0.41	0.03	1.52	0.46	0.41	0.03	1.53
θ_{20}-θ_{29}	1	0.81	0.64	0.11	2.49	0.81	0.63	0.10	2.46
θ_{30}-θ_{33}	2	1.29	0.90	0.25	3.67	1.27	0.89	0.24	3.59
θ_{34}, θ_{35}	3	1.85	1.18	0.43	4.89	1.83	1.19	0.37	4.87
θ_{36}	5	3.15	1.72	0.85	7.39	3.13	1.75	0.81	7.42
μ		−0.70	0.36	−1.51	−0.09	−0.70	0.37	−1.54	−0.08
σ^2		1.21	0.87	0.17	3.45	1.21	0.92	0.17	3.56

algorithm, it yields outstanding results. In this case, the large size of the data set ($m = 36$ as opposed to $m = 9$ for the audit data example) improves the sampling from the full conditional posterior distribution of σ^2, as it increases the number of the degrees of freedom of the chi-square distribution involved in this simulation. We therefore notice that increasing the sample size, has in essence the same effect as introducing some prior knowledge for σ^2, as far as simulation from the full conditional distribution of this parameter is concerned. This is illustrated in Figure 5.15, where the σ^2 simulated values vary in an approximate range of $(0, 15)$, as opposed to a corresponding range of approximately $(0, 700)$ for the audit data example, shown in Figure 5.11. In the latter case, this unusually high variation dropped when we assumed a more informative prior for σ^2. Moreover, our mixture approximation to the full conditional posterior distribution of the Poisson means θ_i, $i = 1, \ldots, m$, will improve as a result of the lower variation of the σ^2 simulated values. This is reflected in the acceptance rate for the Metropolis-Hastings subalgorithm, which now is higher than for the uniform case in the audit data example, lying within the range 93%-99%.

A burn-in period of 4000 simulations was used, as suggested by Figure 5.16, which shows that the value of the Gelman and Rubin R statistic was very close to 1 after the first 8000 iterations of the algorithm. The parameter estimates, reported in Table 5.12 are once again almost identical to the BUGS respective

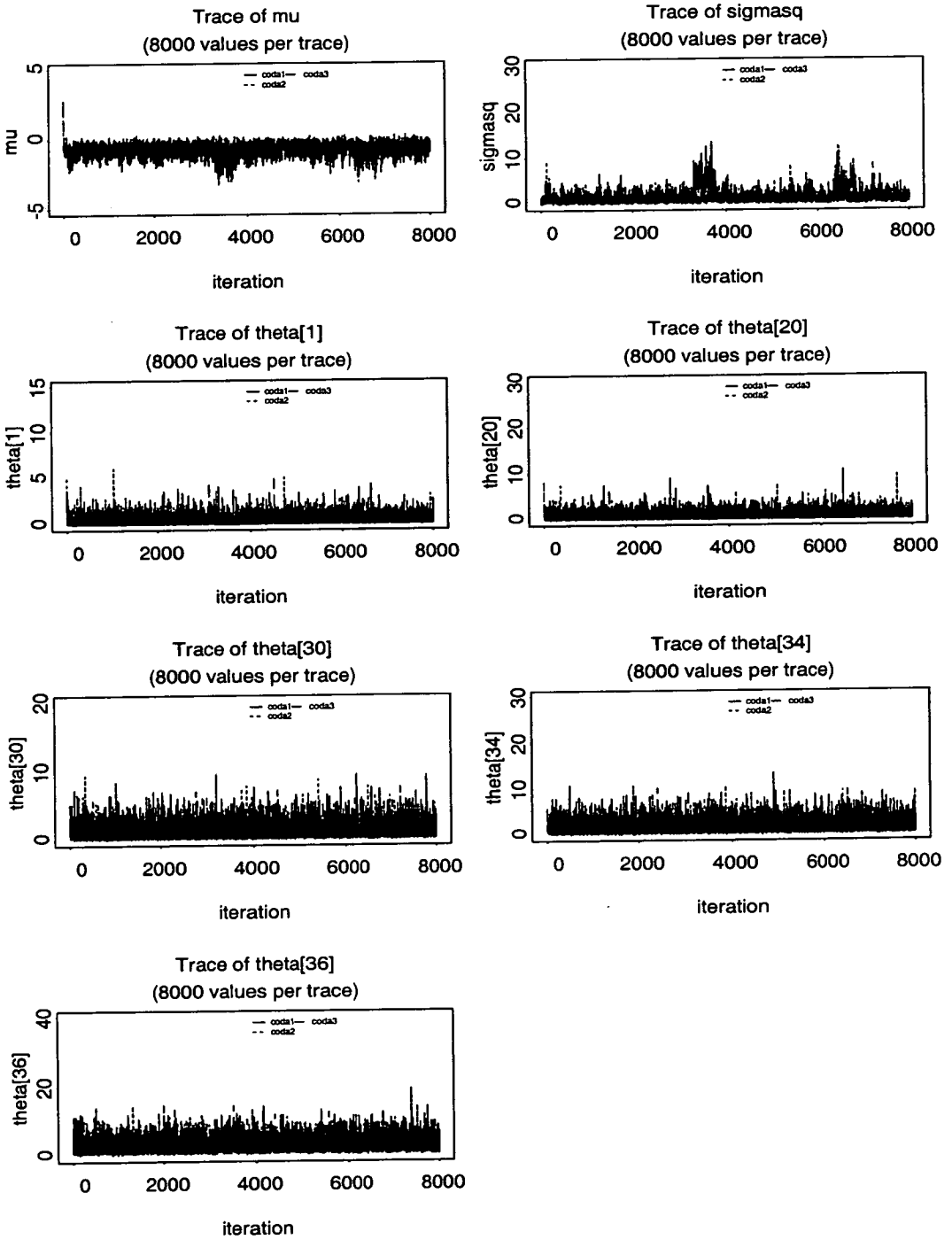

Figure 5.15: *Trace of sampled values from 3 parallel chains for the oilwell discoveries data example parameters. Model (4.1a) is assumed and the mixture proposal is used for the Metropolis-Hastings subalgorithm.*

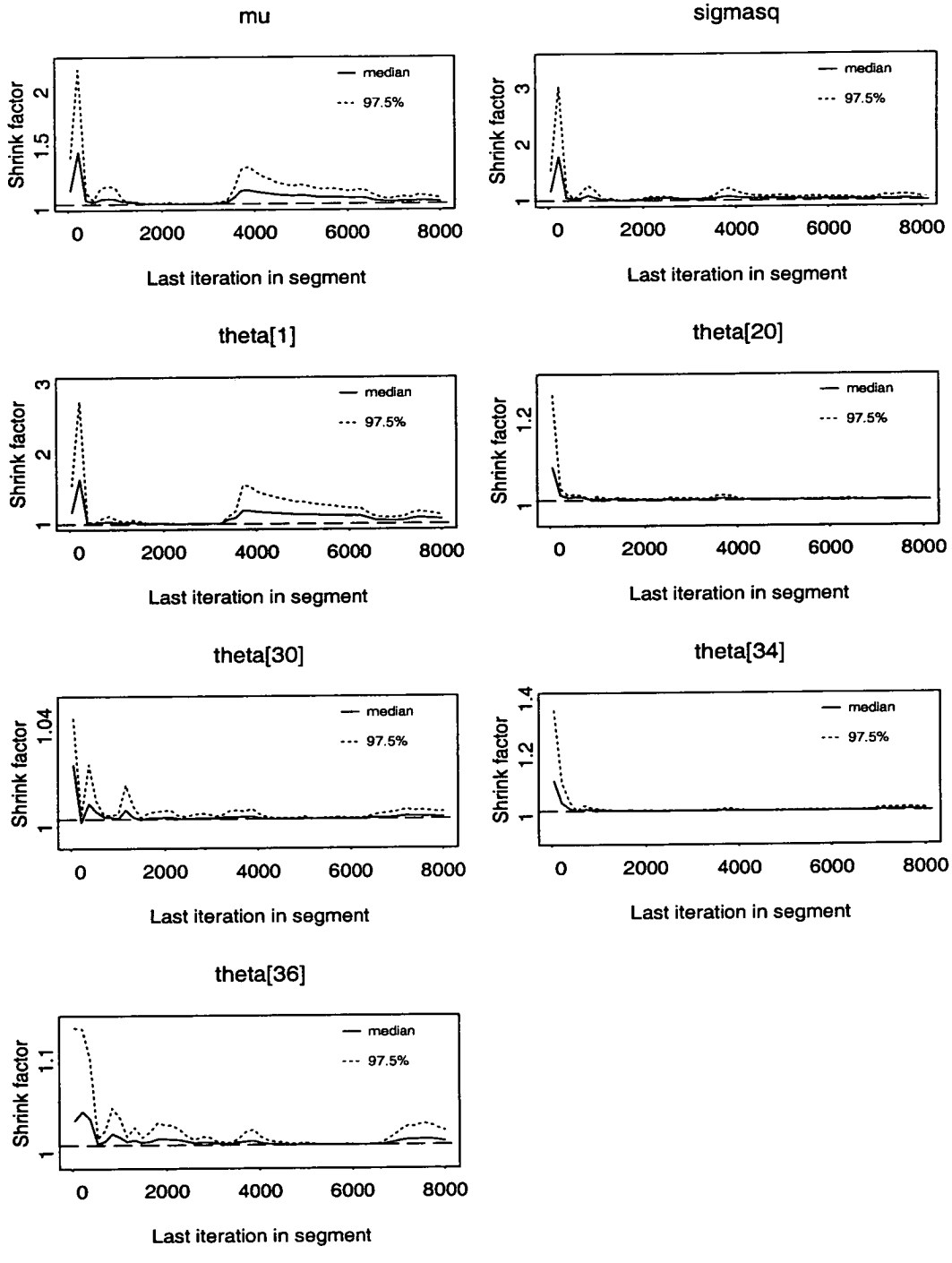

Figure 5.16: *Gelman and Rubin R statistic against iteration number for the oilwell discoveries data example parameters. Model (4.1a) is assumed and the mixture proposal is used for the Metropolis-Hastings subalgorithm.*

results in Table 5.9.

When the $\text{Ga}(y_i + 1, 1)$ proposal distribution is used to draw candidate points for the Metropolis-Hastings subalgorithm, we first notice that the acceptance rate declines to 32%-52%, with the lower rates corresponding to large data values. The general impression that we get from Table 5.12 is that some slight discrepancies with the estimates obtained using the gamma proposal distribution appear, mostly in the cases where low acceptance rates occurred, that is when $y = 2, 3$ or 5. This again suggests that longer chains should be used with this candidate distribution, to account for the effect of higher rejection.

Model (4.1b): Inv-$\chi^2(\nu, \lambda)$ hyperprior on σ^2

The results when a Inv-$\chi^2(10, 0.46)$ prior distribution is assumed for σ^2 are displayed in Table 5.13. In this case the mixture approximation to $p(\theta_i|\mu, \sigma^2, \mathbf{y})$ is further improved, as it benefits from both the large sample size and the prior information that we introduce in the model. This is reflected in an outstandingly high acceptance rate of 97%-98%. Again, convergence is achieved faster under an informative prior distribution for the variance parameter σ^2. Figure 5.17 verifies that the Gelman and Rubin diagnostic is practically equal to 1 after 3000 iterations of the algorithm, and the trace of the σ^2 simulated values in Figure 5.18 exhibits even lower variation than for the uniform σ^2 prior case in Figure 5.15.

Table 5.13: *Metropolis-Hastings within Gibbs estimates of the posterior mean, standard deviation and* (2.5%, 97.5%) *percentiles for the oilwell discoveries data example parameters. Model (4.1b) is assumed, and both mixture and gamma proposals are considered for the Metropolis-Hastings subalgorithm.*

		Mixture proposal				*Gamma proposal*			
		Posterior estimates				*Posterior estimates*			
par.	y_i	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1-θ_{19}	0	0.56	0.39	0.09	1.57	0.56	0.40	0.09	1.59
θ_{20}-θ_{29}	1	0.82	0.54	0.17	2.20	0.81	0.53	0.18	2.14
θ_{30}-θ_{33}	2	1.14	0.71	0.28	2.97	1.13	0.70	0.28	2.97
θ_{34}, θ_{35}	3	1.54	0.90	0.42	3.84	1.52	0.89	0.42	3.80
θ_{36}	5	2.53	1.33	0.77	5.84	2.48	1.36	0.65	5.83
μ		−0.51	0.26	−1.05	−0.05	−0.51	0.25	−1.04	−0.04
σ^2		0.57	0.26	0.25	1.23	0.57	0.26	0.24	1.21

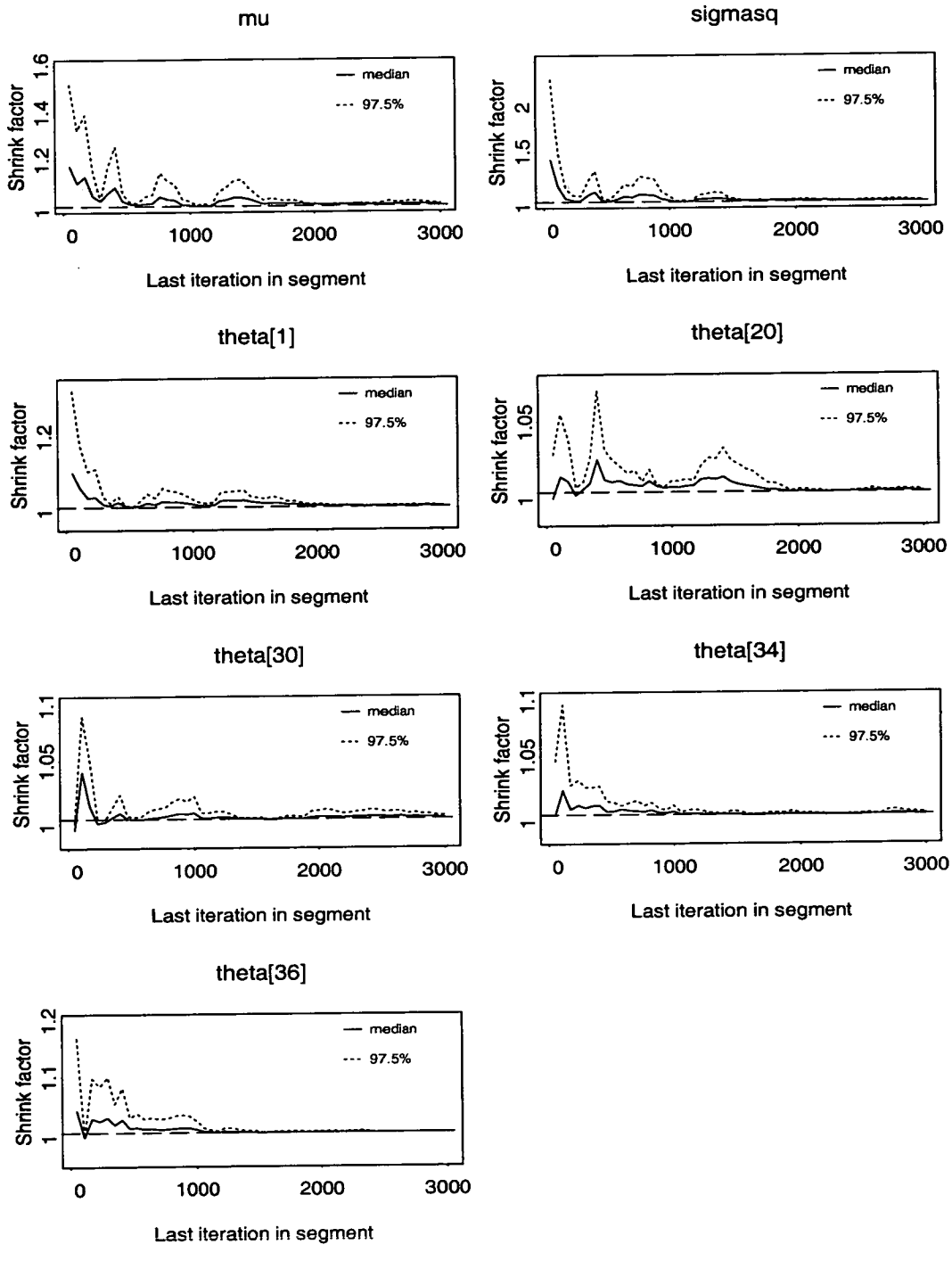

Figure 5.17: *Gelman and Rubin R statistic against iteration number for the oilwell discoveries data example parameters. Model (4.1b) is assumed and the mixture proposal is used for the Metropolis-Hastings subalgorithm.*

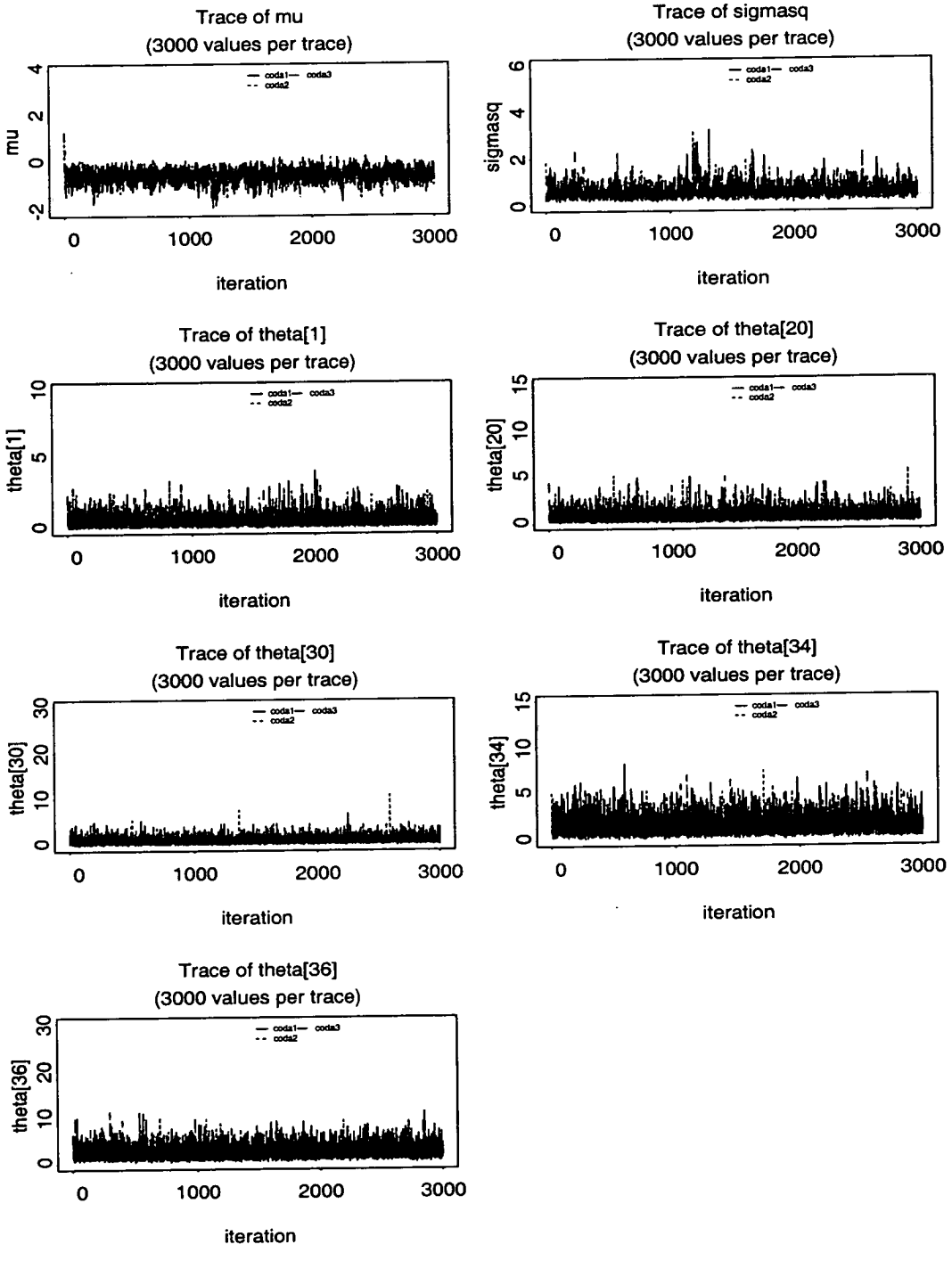

Figure 5.18: *Trace of sampled values from 3 parallel chains for the oilwell data example parameters. Model (4.1b) is assumed and the mixture proposal is used for the Metropolis-Hastings subalgorithm.*

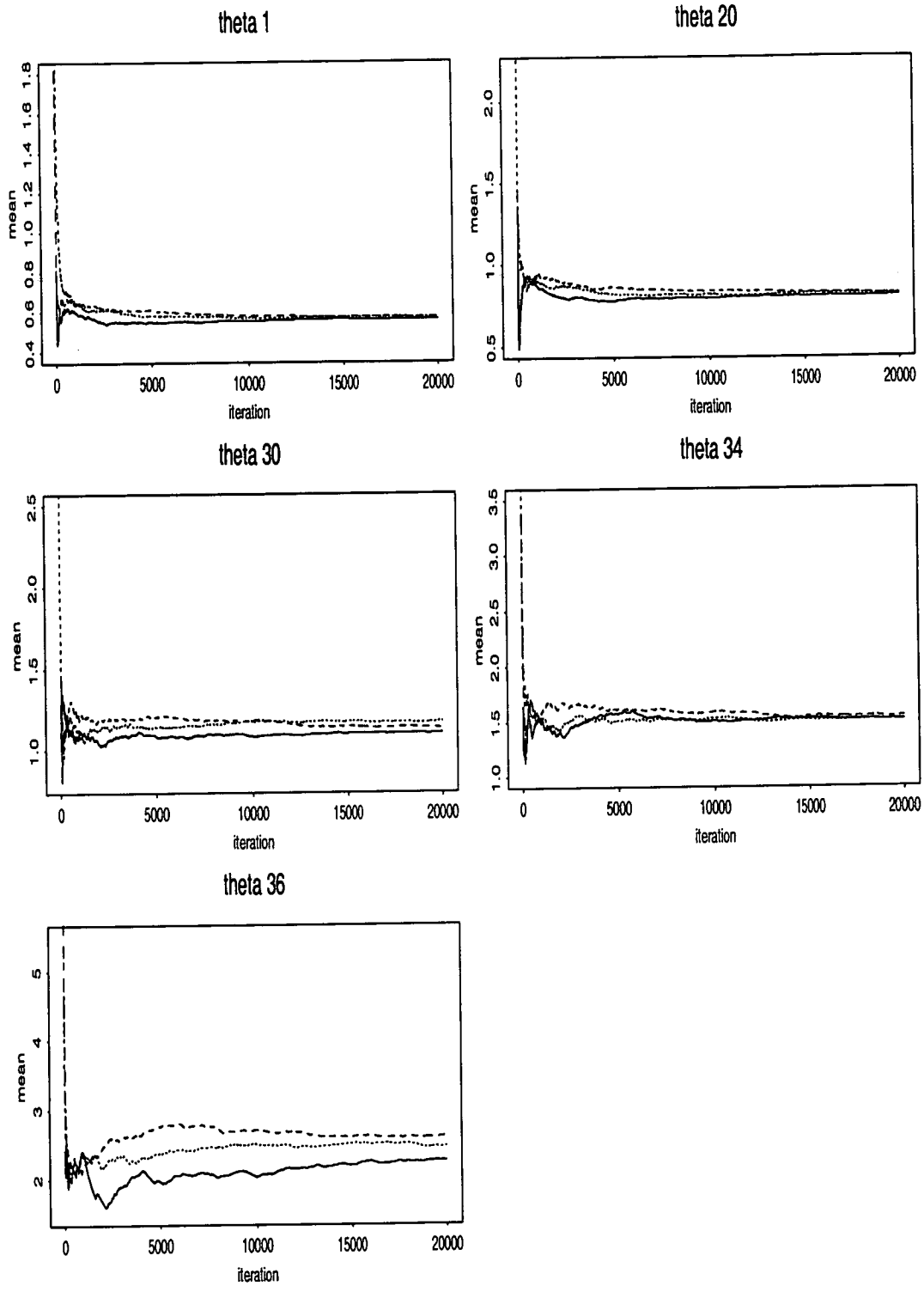

Figure 5.19: *Posterior mean trace for the parameters θ_1, θ_{20}, θ_{30}, θ_{34} and θ_{36} in the oilwell discoveries data example. Model (4.1b) is assumed and the gamma proposal is used in the Metropolis-Hastings subalgorithm. The different lines correspond to the 3 independent Gibbs sampling chains.*

All parameter estimates in Table 5.13 are practically identical to these obtained with BUGS, in Table 5.9. This is also true in the case of σ^2, for which estimates appeared to be slightly inconsistent when the uniform prior distribution was assumed for this parameter. Finally, when we employ a $\text{Ga}(y_i + 1, 1)$ proposal distribution the acceptance rate drops to its lowest levels for all examined cases, that is it lies between 15% and 52%. The worst rates occur when y is away from the origin. It is also in this case that the convergence is delayed. The Gelman and Rubin R statistic for θ_{36} was equal to 1.14 after 6000 iterations, whereas it was equal to 1.00 for the same parameter after 3000 iterations when the mixture proposal was chosen. The estimates, shown in Table 5.13, are inaccurate when the acceptance rate gets lower, that is for the parameters corresponding to relatively large observations. In fact, as illustrated in Figure 5.19, the estimate of the posterior mean for θ_{36}, where the acceptance rate is as low as 15%, does not appear to have converged to the correct value, implying again that a longer chain should be used in such cases.

5.6 Shrinkage behaviour of the hierarchical Bayes estimator

As discussed in the preceding chapters, the posterior mean of the Poisson parameters θ_i, $i = 1, \ldots, m$, can be viewed as a shrinkage estimator. In the case of the first stage conjugate Poisson/Gamma structure, the posterior mean can be exactly expressed as a linear estimator, adjusting the unbiased estimates Y_i, $i = 1, \ldots, m$, towards the prior mean of the parameters of interest. In Subsection 3.3.2 and in Section 4.3 we assumed similar linear approximations to the posterior mean for the Poisson/log-normal model. In the EB context, the form of the linear approximation given in (3.24), implies that the estimator shrinks the observed values towards the sample mean of the data. However, both the nonlinear importance sampling estimator, described in Subsection 3.4.1, and the approximate hierarchical Bayes analysis in Section 4.3, suggested that the posterior mean adjusts the unbiased estimators of the Poisson parameters θ_i, $i = 1, \ldots, m$, towards a central point, which is not given by the sample mean for a finite sample size.

We will now consider the two artificial data sets, which were analysed in Section 3.8 under an EB framework, to investigate the shrinking behaviour of the Bayesian estimates of the Poisson means when a full hierarchical analysis is followed. We consider the Poisson/log-normal structure described in (4.1a), that is assuming vague prior distributions for both hyperparameters μ and σ^2. The exact full hierarchical analysis is conducted employing the Metropolis-Hastings within

Table 5.14: *Hierarchical Bayes estimates of all the model parameters in the first simulated data example. The sample mean is equal to* $\bar{y} = 5.22$ *and the sample variance is* $s^2 = 13.94$.

		Posterior estimates	
par.	y	*mean*	*s.d.*
θ_1	1	2.22	1.33
θ_2,θ_3	2	2.83	1.47
θ_4	3	3.50	1.60
θ_5,θ_6	5	4.91	1.97
θ_7	8	7.21	2.46
θ_8	9	8.04	2.64
θ_9	12	10.50	3.21
μ		1.42	0.40
σ^2		1.11	1.36

Gibbs method with the log-normal/gamma proposal mixture distribution. Initialising the algorithm at different starting values, three independent chains of length $N = 2 \times 10^4$ each, were ran for the two examples. In both cases, the Gelman And Rubin R diagnostic was virtually equal to 1 after 3000 simulations in each chain, and therefore a burn-in period of 1500 iterations was utilised to allow convergence to the target posterior distribution. The posterior estimates are given in Tables 5.14 and 5.15.

As with the nonlinear EB importance sampling estimator, it is shown from the posterior estimate of θ_5 or θ_6 for $y = 5$ in Table 5.14, and from the estimate of θ_9 for $y = 9$ in Table 5.15, that the hierarchical Bayes method does not shrink the MLE towards the sample mean, that being $\bar{y} = 5.22$ and $\bar{y} = 9.27$ for the first and second simulated data set respectively. Hence, the hierarchical Bayes posterior mean contradicts the shrinking behaviour of previously proposed EB or frequentist shrinkage estimators, suggesting that the shrinking direction should be determined from the entire information included in the model, rather than be chosen exclusively based on the data or as a fixed point.

Table 5.15: *Hierarchical Bayes estimates of all the model parameters in the second simulated data example. The sample mean is equal to* $\bar{y} = 9.27$ *and the sample variance is* $s^2 = 24.64$.

		Posterior estimates	
par.	y	*mean*	*s.d.*
θ_1,θ_2	4	5.98	1.98
θ_3	5	6.53	2.06
θ_4,θ_5	6	7.12	2.14
θ_6,θ_7	7	7.71	2.23
θ_8	8	8.31	2.33
θ_9	9	8.94	2.44
θ_{10},θ_{11}	10	9.61	2.57
θ_{12},θ_{13}	12	10.90	2.81
θ_{14}	17	14.40	3.49
θ_{15}	22	18.20	4.20
μ		2.14	0.16
σ^2		0.26	0.18

5.7 Summary and conclusions

In this chapter we have discussed stochastic simulation approaches to the full hierarchical Bayesian analysis of the the Poisson/log-normal model.

We first attempted the use of importance sampling employing an importance density given by a slightly adjusted form of the γ likelihood function. The method seemed to provide reliable results only in limited cases, and after an unusually large number of Monte Carlo simulations. Unbounded weights and high standard errors of simulation suggested that the proposed importance density is not suitable in many situations, including the cases when little prior knowledge is assumed in the second level of the prior specification, or when many zero counts are observed.

MCMC techniques can also be employed. The Gibbs sampler was initially considered. To overcome the problem of sampling from the nonclosed form of the full conditional posterior distribution of the Poisson means θ_i, we developed a log-normal/gamma mixture approximation to the density of interest, based on matching the first 3 moments of the exact and the approximate distributions.

The posterior moments were obtained with a combination of a technique based on a discretisation of the prior distribution and use of Bayes' theorem to obtain a discrete approximation to the posterior distribution, and an entropy distance minimisation method. The Gibbs sampler results were close to the correct estimates in all the cases that we examined, especially when a large sample, or strong second stage prior information was available. However, some occurred discrepancies, due to possible inaccuracies in the approximation of the moments of $p(\theta_i|\mu, \sigma^2, \mathbf{y})$, suggest that one might choose to follow a different strategy, leading to the exact estimation.

We therefore adopted a hybrid MCMC approach, namely the Metropolis-Hastings within Gibbs technique. The developed log-normal/gamma mixture approximation to $p(\theta_i|\mu, \sigma^2, \mathbf{y})$ was employed as the proposal distribution for the Metropolis-Hastings subalgorithm, offering an outstandingly high acceptance rate due to its resemblance to the exact conditional distribution. We also considered a $\text{Ga}(y_i + 1, 1)$ proposal density. Both candidate distributions provided the correct estimates. The final choice between one of the two seems to involve a trade-off between evaluation effort and mixing speed, as the mixture proposal demanded more computing time to evaluate, but offered higher acceptance rate and faster mixing.

We also investigated the shrinking behaviour of the hierarchical Bayes estimator, in an attempt to compare it to previous suggestions regarding shrinkage towards the origin, the minimum observation, the sample mean, the median, some prior guess etc. Our experimentation with simulated data sets verified the conclusions of Chapter 3 for the EB estimation, and those of Chapter 4 for the approximate hierarchical Bayes solution, that the posterior mean shrinks the unbiased MLE towards a point lying between the minimum observation in the data and the sample mean.

Chapter 6

Hierarchical Bayes frequency properties and applications

6.1 Frequency properties of the hierarchical Bayes estimator

In Chapter 5 a full hierarchical Bayes analysis was presented for the Poisson/log-normal model. The estimation of the model parameters was achieved through MCMC procedures. We now want to assess the frequency properties of the hierarchical Bayes estimator of θ_i, $i = 1, \ldots, m$, using the criterion of average risk. This was introduced in Definition 3.1, and was employed in Section 3.9 to examine the frequency properties of the empirical Bayes (EB) estimators developed in Chapter 3, that is the importance sampling estimator and the best linear predictor (BLP), and compare them to some of the methods suggested in the literature.

6.1.1 Modelling issues and frequency properties

Until now, in the full Bayesian analysis we have used two different hierarchical structures for the Poisson/log-normal model. The first, in (4.1a), assumes a $U(0, \infty)$ hyperprior distribution for the variance σ^2 of $\gamma_i = \log(\theta_i)$, $i = 1, \ldots, m$, at the second sage of the prior setting, while in the second structure, in (4.1b), we replace the vague uniform distribution with the conjugate Inv-$\chi^2(\nu, \lambda)$ distribution. In both cases, a vague flat hyperprior distribution is assumed for the mean parameter μ. Our aim is to investigate the average risk properties of the hierarchical Bayes estimator under both prior structures, as these can represent two opposite situations regarding the available prior information. The uniform distribution is chosen either to show prior ignorance for σ^2, or in an effort to avoid any preference on particular values of the parameter. On the other hand, the scaled inverse chi-square distribution, while it may be tuned to provide rel-

atively vague prior knowledge, it is often used to indicate the presence of some degree of information in that level of the hierarchy.

When a Inv-$\chi^2(\nu, \lambda)$ prior distribution for σ^2 is used, we choose the parameters ν and λ so that relatively strong prior knowledge is assumed. We do that in the same way as for the examples in Chapter 5. We set $\nu = 10$ and then we choose λ in such a way that the expectation of the Inv-$\chi^2(\nu, \lambda)$ distribution matches a data-determined estimate of σ^2. We notice that, as described in Section 3.5 the method of moments for σ^2 gives

$$\widehat{\sigma}^2 = \log\left\{1 + \frac{(s^2 - \bar{y})_+}{\bar{y}^2}\right\},$$

and then, setting this value equal to the expected value of the Inv-$\chi^2(\nu, \lambda)$ distribution we get

$$\mathrm{E}(\sigma^2) = \widehat{\sigma}^2 \Rightarrow \frac{\nu}{\nu - 2}\,\lambda = \widehat{\sigma}^2$$

from which we obtain

$$\lambda = \frac{\nu - 2}{\nu}\,\widehat{\sigma}^2.$$

We also notice that the variance of the prior distribution can then be given by

$$\mathrm{var}(\sigma^2) = \frac{2\nu^2}{(\nu - 2)^2(\nu - 4)}\,\lambda^2.$$

The above procedure gives a hyperprior distribution which is closely concentrated around the data-estimated value of the parameter σ^2. For instance, in the audit data example the chosen values $\nu = 10$, $\lambda = 0.45$, give a hyperprior distribution for σ^2 which is concentrated at a data estimated mean equal to 0.56 with a variance of 0.05. We note that $\widehat{\sigma}^2 = 0.56$ is the σ^2 value employed in the EB methods for the same example in Chapter 3. Thus, by using such an informative prior for σ^2, we expect the posterior mean estimates produced with the hierarchical Bayes analysis to be similar to those of the EB methods.

6.1.2 MCMC methods for the frequency simulations

In Chapter 5 we obtained the hierarchical Bayes analysis of the Poisson/log-normal model using the the Gibbs sampling method developed in Section 5.4 and the hybrid MCMC strategy of Section 5.5. The former relies on a log-normal/gamma mixture approximation to simulate from the nonclosed form of the full conditional posterior distribution of θ_i, $i = 1, \ldots, m$, while the latter addresses the same problem employing a Metropolis-Hastings step, embedded within the Gibbs sampling cycles.

We assess the frequency properties of the hierarchical Bayes estimator under both approaches. The Gibbs sampling algorithm, was shown through the examples of the preceding chapter to provide a good approximation to the posterior distribution of the model parameters. The simulation study will demonstrate that it also produces an estimator with good average risk properties, especially in comparison to the usual unbiased estimator or other frequentist methods. The Metropolis-Hastings within Gibbs hybrid method will be used to demonstrate the frequency properties of the estimator resulting from an exact hierarchical Bayes analysis of the model under consideration. The log-normal/gamma mixture distribution was preferred to the $\mathrm{Ga}(y_i + 1, 1)$ as the proposal distribution for the Metropolis step of the algorithm. This was due to the mixing efficiency advantage of the former, in anticipation of faster convergence, which was crucial given the restrictions dictated by the computationally intensive nature of the simulation study.

6.1.3 MCMC implementation and starting values

The estimation of the average risk of the hierarchical Bayes estimator through a simulation procedure requires the MCMC algorithm to be repeated for a large number of times, to allow the evaluation of (3.32). This imposes some restrictions regarding the length of the chain of the algorithm. The latter is related to one of the main concerns with the implementation of MCMC methods, namely the issue of the convergence of the Markov chain to the distribution of interest. The chain must be sampled for as long as it is necessary for the simulated values to be drawn from the correct distribution. The sufficient length of the MCMC chain can vary according to the problem under consideration, and therefore a case-to-case assessment of convergence must be made to allow a decision regarding the length of the chain. When the Gibbs sampling methods of Chapter 5 were applied to specific examples, the convergence was examined by means of visual inspection of certain characteristics of the resulting distribution, and also using the Gelman and Rubin (1992) R statistic. However, these convergence diagnostic methods rely on running multiple simultaneous MCMC chains and involve an interactive procedure. It is therefore difficult to determine an all-purpose rule for the appropriate length of the Markov chain when a simulation study involving a large number of different samples is to be carried. The parallel implementation of multiple chains for the assessment of convergence and better mixing of the algorithm, was highly time consuming during the repeated iterations of the frequency properties study.

To tackle this problem, we only ran a single chain for each MCMC evaluation of the hierarchical Bayes estimator. Then, one possible way to facilitate the con-

vergence of the algorithm was to choose good starting values, that is to initialise the Markov chain at values which would be close to the posterior estimates of the model parameters. Hence, we employed the EB estimates of the parameters of the Poisson/log-normal model, derived in Chapter 3. The BLP for the Poisson parameters θ_i, $i = 1, \ldots, m$, in (3.24), and the moment estimates (3.23) of the hyperparameters μ and σ^2 provide approximations to the posterior estimates, while their computation is remarkably simple and fast. Then, a single chain of length $N = 5000$ was ran and the burn-in period was kept as low as possible for both the Gibbs sampling and the Metropolis-Hastings within Gibbs methods. Once again, we stress here that these values do not guarantee convergence of the algorithm, but were empirically chosen on a trial and error basis. As also demonstrated in the examples of Chapter 5, our experimentation suggested that the convergence issue can be considerably complicated when the Poisson means are close to zero and little or no prior knowledge is included in the model.

6.1.4 Characteristics of the average risk simulation study

As in Chapter 3, we are interested in investigating the average risk behaviour of the hierarchical Bayes estimates under different loss functions and choosing various parameter settings, representing a wide range of the parameter space of $\boldsymbol{\theta}$. Thus, we consider the squared error loss, the absolute error loss and the maximum component loss functions given in (3.25), (3.26) and (3.27) respectively, and the true mean and variance combinations displayed in Table 3.7. The average risk was estimated using (3.32), and the results are again given terms of the relative average risk improvement (RARI) in (3.33), that is as the relative savings in average risk when the considered method is compared to the usual maximum likelihood estimator.

Finally, a total number of $N = 10^4$ Monte Carlo simulations were employed for the estimation of the average risk, involving 100 simulated data samples, for the evaluation of the frequentist risk, nested within 100 simulations of the $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^{\mathrm{T}}$ vector, to allow averaging over the prior distribution. The algorithm for the C computer program used for the simulation study is the same as the one in Figure 3.1, for the EB case. However, the iterative nature of the MCMC computations adds a considerably heavier computing time cost to the whole algorithm.

Table 6.1: *Percentage of relative improvement in average risk (3.33) when the hierarchical Bayes estimator obtained with the Metropolis-Hastings within Gibbs algorithm is compared to the MLE. Model (4.1a) is assumed, and the results from the 9 settings for the true θ_i, $i = 1, \ldots, m$, values and the 9 considered loss functions are reported.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
SEL_0	41.7	32.9	25.1	52.3	40.7	23.4	51.7	40.1	26.8
SEL_1	43.1	35.1	28.7	52.9	41.7	24.8	51.9	41.1	27.8
SEL_2	41.5	26.5	4.2	52.5	39.0	14.0	51.4	39.7	23.6
AEL_0	28.2	23.3	19.4	32.7	24.9	14.6	31.8	24.0	15.8
$AEL_{\frac{1}{2}}$	29.7	24.9	21.1	33.0	25.2	15.0	31.9	24.2	16.0
AEL_1	30.8	24.6	17.8	33.1	25.0	13.7	31.9	24.2	15.6
$MAXSEL_0$	39.3	29.5	21.0	51.3	38.6	19.5	51.0	38.8	24.0
$MAXSEL_1$	39.4	32.4	26.7	51.9	39.6	21.5	51.3	40.4	26.3
$MAXSEL_2$	34.0	19.2	−5.4	50.2	31.7	−3.3	49.8	36.0	15.0

6.1.5 Average risk simulation results

The relative improvement in average risk of the hierarchical Bayes estimator when compared to the usual unbiased estimator is given in Tables 6.1, 6.3, 6.4 and 6.5. Each of these tables exhibits the RARI for the 9 different specifications for the mean $E(\theta_i)$ and the variance $var(\theta_i)$ of the true values of the Poisson means θ_i, $i = 1, \ldots, m$, and the 9 considered loss functions given in (3.25), (3.26) and (3.27). Each table corresponds to one of the two MCMC methods that we have presented, that is the approximate Gibbs sampling approach and the hybrid Metropolis-Hastings within Gibbs algorithm, when either model (4.1a) or model (4.1b) is assumed.

Model (4.1a): Uniform hyperprior on σ^2

We first look into the case where we assume that the variance parameter σ^2 has a vague $U(0, \infty)$ hyperprior distribution under model (4.1a). The RARI results of the Metropolis-Hastings within Gibbs method in Table 6.1 demonstrate that the estimator resulting from the full hierarchical Bayes analysis, produces remark-

Table 6.2: *Summary of the relative improvement in average risk (RARI) results of the hierarchical Bayes estimator obtained with the Metropolis-Hastings within Gibbs method under both models (4.1a) and (4.1b), when compared to the MLE. A total number of* 81 *cases were examined.*

% *RARI*	< 0	$0 - 20$	$20 - 40$	$40 - 60$	> 60
Model (4.1a)	2	13	46	20	0
Model (4.1b)	7	10	32	22	10

able savings over the average risk of the MLE under almost all the considered loss functions and parameter settings for $\boldsymbol{\theta}$. As expected, the improvement declines as the true variance $\text{var}(\theta_i)$ increases, since in this case the estimates obtained with the hierarchical Bayes method are less adjusted towards the prior assumption and closer to the unbiased estimates. The method results in worse frequency properties than the MLE only in two cases, when the weighted maximum component loss function $\text{MAXSEL}_2 = \max_{1 \le i \le m} \left\{ \frac{(\delta_i - \theta_i)^2}{\theta_i^2} \right\}$ is considered, and the variance of θ_i, $i = 1, \ldots, m$, is large.

Table 6.2 provides a summary of the relative average risk improvement of the hierarchical Bayes estimator, compared to the MLE. As mentioned earlier, when model (4.1a) is assumed, the hierarchical Bayes method fails only twice, out of 81 cases examined, to improve the average risk of the unbiased estimator. We can compare the overall performance of the hierarchical Bayes method to that of the nonlinear EB importance sampling estimator, referring to Tables 6.2 and 3.8. Table 3.8 reveals that the EB method is worse than the MLE in terms of average risk in 6 cases, as compared to the 2 cases for the hierarchical Bayes estimator. On the other hand, the latter did not give an improvement greater than 60% in any of the examined cases, whereas the EB importance sampling estimator improved the MLE in such a high percentage 12 times. As Table 6.1 shows, these findings reflect the good performance of the hierarchical Bayes method, not only when the Poisson parameters θ_i, $i = 1, \ldots, m$, are close to their mean, but also in the presence of larger variability among them. While the EB method, due to ignoring the uncertainty in the hyperparameters, yields outstanding savings when $\text{var}(\theta_i)$ is small, thus resulting in improvement more than 60% in theses cases, the hierarchical Bayes analysis provides good frequency properties also under more variation in the parameters, improving the average risk of the MLE in most of the cases where $\text{var}(\theta_i)$ is large.

We also assessed the average risk properties of the hierarchical Bayes estimator

Table 6.3: *Percentage of relative improvement in average risk (3.33) when the hierarchical Bayes estimator obtained with the approximate Gibbs sampling algorithm is compared to the MLE. Model (4.1a) is assumed, and the results from the 9 settings for the true θ_i, $i = 1, \ldots, m$, values and the 9 considered loss functions are reported.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
SEL_0	51.5	43.9	32.9	50.3	38.6	26.4	52.1	39.8	26.4
SEL_1	53.3	45.7	31.1	51.1	40.4	28.6	52.4	40.5	27.0
SEL_2	48.3	26.6	−68.4	50.7	37.9	21.5	52.2	39.0	21.7
AEL_0	36.4	32.1	25.4	31.2	23.6	16.5	32.1	23.4	15.4
$AEL_{\frac{1}{2}}$	38.2	33.2	22.5	31.6	24.2	17.2	32.2	23.6	15.5
AEL_1	38.6	30.1	4.9	31.8	24.1	16.6	32.3	23.4	14.8
$MAXSEL_0$	48.1	40.1	26.3	48.9	35.6	23.2	51.7	38.6	23.8
$MAXSEL_1$	49.6	43.7	31.6	49.9	38.2	25.5	52.2	40.0	25.2
$MAXSEL_2$	39.5	16.4	−96.9	47.8	30.6	6.4	51.2	35.3	12.1

resulting from the method presented in Section 5.4, that is the Gibbs sampling method employing the log-normal/gamma mixture approximation for sampling from the full conditional distribution of θ_i, $i = 1, \ldots, m$. As shown in Table 6.3, the Gibbs sampling estimator again outperforms the MLE in all but two of the 81 considered cases. The relative improvement is as high as that of the exact Metropolis-Hastings within Gibbs estimator, and follows similar patterns. There is a reasonable agreement in the RARI results of the exact and the approximate method, in Tables 6.1 and 6.3 respectively, except in the case where $E(\theta_i) = 1.0$. We believe that these discrepancies can be due to the inaccuracies occurred with the approximate algorithm when the data include many zero values, as would the case be when the true mean of the Poisson parameters θ_i, $i = 1, \ldots, m$, is close to the origin.

Model (4.1b): Inv-$\chi^2(\nu, \lambda)$ hyperprior on σ^2

As mentioned before, the uniform hyperprior distribution on σ^2, that we assumed in the hierarchical Bayes setting for the preceding paragraph, implies relative prior

Table 6.4: *Percentage of relative improvement in average risk (3.33) when the hierarchical Bayes estimator obtained with the Metropolis-Hastings within Gibbs algorithm is compared to the MLE. Model (4.1b) is assumed, and the results from the 9 settings for the true θ_i, $i = 1, \ldots, m$, values and the 9 considered loss functions are reported.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
SEL_0	63.3	53.5	44.4	61.4	45.0	27.7	60.2	42.4	24.6
SEL_1	60.4	44.5	21.3	61.5	45.0	24.0	60.2	43.0	24.5
SEL_2	40.8	−7.3	−155.3	59.2	36.8	0.5	58.7	39.2	14.0
AEL_0	42.5	34.6	28.6	39.1	27.7	16.0	38.0	25.3	14.3
$AEL_{\frac{1}{2}}$	41.1	30.0	18.4	39.1	27.4	14.6	38.0	25.4	14.1
AEL_1	36.6	18.2	−11.3	38.6	25.8	10.4	37.7	24.7	12.4
$MAXSEL_0$	62.6	51.7	41.6	60.6	42.0	25.3	59.5	39.7	21.2
$MAXSEL_1$	63.1	50.0	22.0	61.4	43.4	21.9	59.7	42.1	23.2
$MAXSEL_2$	30.6	−21.1	−210.6	55.7	24.0	−27.6	55.8	32.2	−1.5

ignorance at the lowest stage of the hierarchy. However, using a scaled inverse chi-square hyperprior distribution on σ^2, we can include in the model certain prior information for the variance hyperparameter. We assume a Inv-$\chi^2(\nu, \lambda)$ prior distribution, determining ν and λ as described in Subsection 6.1.1. In doing so, we assume that a-priori, σ^2 is distributed around a data-determined mean with relatively small variation. Therefore, we expect the estimator obtained from this hierarchical Bayes analysis to exhibit smaller posterior variability than before, and hence to possess excellent frequency properties when the parameters of interest θ_i, $i = 1, \ldots, m$, are close to each other, but not when they lie in a wide range.

Table 6.4 contains the relative savings in average risk when the hierarchical Bayes estimator, obtained with the Metropolis-Hastings within Gibbs algorithm for model (4.1b), is compared to the MLE. The comparison of these RARI results to those given in Table 6.1, verifies that the assumption of strong hyperprior information provides estimators with superior frequency properties when $var(\theta_i)$ is small. However, the average risk performance of the method derived under vague priors, is similar or better than that of the method assuming a strongly

Table 6.5: *Percentage of relative improvement in average risk (3.33) when the hierarchical Bayes estimator obtained with the approximate Gibbs sampling algorithm is compared to the MLE. Model (4.1b) is assumed, and the results from the 9 settings for the true θ_i, $i = 1, \ldots, m$, values and the 9 considered loss functions are reported.*

$E(\theta_i)$	1.0			5.0			10.0		
$var(\theta_i)$	0.5	1.0	2.0	2.5	5.0	10.0	5.0	10.0	20.0
SEL_0	59.7	49.5	43.9	55.9	45.3	31.0	56.2	45.1	32.0
SEL_1	61.2	48.2	32.4	56.3	45.6	30.6	56.4	45.8	31.3
SEL_2	52.7	15.4	−80.5	55.1	40.2	9.2	56.1	44.2	23.3
AEL_0	41.6	34.3	29.2	35.1	27.5	19.1	34.9	26.9	18.4
$AEL_{\frac{1}{2}}$	42.6	33.1	22.4	35.4	27.6	18.8	35.1	27.1	18.0
AEL_1	41.6	26.2	−0.7	35.3	26.8	15.9	35.1	26.9	16.6
$MAXSEL_0$	56.6	46.0	39.6	54.8	44.0	26.4	56	43.6	30.2
$MAXSEL_1$	61.2	50.2	38.2	55.1	44.3	26.2	56.2	44.9	29.5
$MAXSEL_2$	46.1	2.9	−95.7	51.5	30.4	−21.9	54.6	40.3	10.0

informative Inv-$\chi^2(\nu, \lambda)$ prior, when the θ_i, $i = 1, \ldots, m$, are less concentrated around their mean. This is also demonstrated in Table 6.2, which shows that the hierarchical Bayes method, when an informative prior on σ^2 is assumed, fails to improve the average risk of the MLE in 7 out of the 81 examined cases, while it yields outstanding relative savings of over 60% in 10 cases. Reference to Table 6.4 confirms that the poor frequency performance corresponds to the cases where $var(\theta_i)$ is large, whereas the highest savings are observed when θ_i, $i = 1, \ldots, m$, are very closely distributed around their mean. We note here that this average risk behaviour is very similar to that of the EB methods, which are discussed in Chapter 3 and summarised in Table 3.8. Similarly to the EB estimation, when informative priors are utilised in the model, the hierarchical Bayes analysis implies strong belief in the prior assumptions, and thus is expected to provide excellent frequency results in the parameter region where the prior distribution attaches high probability, at the expense of paying a heavier risk cost under any prior misspecification.

As in the previous paragraph, we also examined the average risk properties of

the estimator resulting from the approximate Gibbs algorithm, when model (4.1b) is assumed. The RARI results in Table 6.5 demonstrate that the method provides remarkable improvement to the MLE as far as the average risk is concerned, giving similar results to those obtained with the exact hierarchical Bayes analysis.

6.2 Model extensions and applications

In the preceding chapters we have considered the analysis of hierarchical Poisson models, where the parameters $\theta_i,\ i = 1, \ldots, m$, represent the Poisson means over equal time or space intervals. Moreover, the Poisson means are given in an exchangeable form, in the sense that they do not depend on any explanatory variables in a regression structure. In the present section we will demonstrate that the methods that we proposed for the hierarchical Bayes analysis of this model, can be easily adjusted to apply to more complicated structures.

6.2.1 Analysis of event rates

When the observed events are counted over time intervals or space areas of unequal sizes, one can consider the rate of occurrence of the events per time or space unit. If the counts $y_1, y_2, \ldots, y_m$, are observed over the time, say, intervals $E_i,\ i = 1, \ldots, m$, with assumed unobserved rates $\theta_i,\ i = 1, \ldots, m$, then we can consider that given the parameters θ_i, the random variables Y_i follow independent Poisson distributions with respective means $\theta_i E_i$ for $i = 1, \ldots, m$. The quantities $E_i,\ i = 1, \ldots, m$, are often called exposure times, indicating the time period for which the experimental units are exposed to a process, condition, treatment, etc. They may also be referred to as denominators, a term which is motivated by the Poisson approximation to the binomial distribution in applications where y_i out of E_i subjects respond to a given treatment.

Assuming a log-normal prior distribution for the event rates, and vague hyperprior distributions for the log-normal parameters, the above situation can be expressed as an exchangeable hierarchical structure which is slightly different from model (4.1a). We can write

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i E_i) \\ \gamma_i &= \log(\theta_i) \\ \gamma_i|\mu, \sigma^2 &\overset{iid}{\sim} N(\mu, \sigma^2) \\ \pi(\mu) &\propto 1 \\ \pi(\sigma^2) &\propto 1 \end{aligned} \tag{6.1}$$

where $i = 1, \ldots, m$. According to this model the joint posterior distribution of $\boldsymbol{\gamma}$, μ and σ^2 is given by

$$p(\boldsymbol{\gamma}, \mu, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}m} \exp\left[\sum_{i=1}^{m}\left\{\gamma_i y_i - E_i e^{\gamma_i} - \frac{1}{2}\sigma^{-2}(\gamma_i - \mu)^2\right\}\right],$$

which implies that while the full conditional distributions of μ and σ^2 remain the same as for model (4.1a), the full conditional of γ_i, $i = 1, \ldots, m$, must be altered, and therefore our log-normal/gamma mixture approximation, as developed in Chapter 5, is no longer valid. However, model (6.1) may be equivalently written as

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\mu_i) \\ \gamma_i &= \log(\mu_i) \\ \gamma_i|\alpha_i, \sigma^2 &\overset{ind}{\sim} N(\alpha_i, \sigma^2) \\ \alpha_i &= \log E_i + \mu \\ \pi(\mu) &\propto 1 \\ \pi(\sigma^2) &\propto 1 \end{aligned} \tag{6.2}$$

with $\mu_i = \theta_i E_i$, for $i = 1, \ldots, m$. The formulation in (6.2) is the same as that of model (4.1a), with the mean of the prior normal distribution replaced by α_i. Hence, the full conditional distributions under model (4.1a) are slightly changed to account for the exposures E_i, and can be written as

$$\mu \mid \boldsymbol{\gamma}, \sigma^2, \mathbf{y} \sim N\left(\frac{\sum_i^m(\gamma_i - \log E_i)}{m}, \frac{\sigma^2}{m}\right)$$

$$\frac{\sum_{i=1}^m(\gamma_i - \alpha_i)^2}{\sigma^2} \;\Big|\; \boldsymbol{\gamma}, \mu, \mathbf{y} \sim \chi^2_{m-2}$$

$$p(\gamma_i|\mu, \sigma^2, \mathbf{y}) \propto \exp\left\{\gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\,\sigma^{-2}(\gamma_i - \alpha_i)^2\right\}, \qquad i = 1 \ldots, m,$$

where $\alpha_i = \log E_i + \mu$. To sample from the full conditional distribution of γ_i we can now employ the log-normal/gamma mixture approximation developed in Chapter 5, substituting μ by α_i. The simulation procedure will provide estimates for the Poisson means μ_i, $i = 1, \ldots, m$, and therefore one should divide these values by the exposures E_i to obtain the Poisson rates θ_i, $i = 1, \ldots, m$.

Table 6.6: *Pump failure data (Worledge* et al., *1982). The exposure times are given in thousands of hours.*

Pump i	Failures y_i	Times E_i
1	5	94.320
2	1	15.720
3	5	62.880
4	14	125.760
5	3	5.240
6	19	31.440
7	1	1.048
8	1	1.048
9	4	2.096
10	22	10.480

6.2.2 Example: Pump failure data

We will illustrate the application of the methods presented in Chapter 5 to the analysis of model (6.2), using the pump failure data which can be found in Worledge *et al.* (1982). The data represent the number of failures of pumps at a nuclear power plant. The failures are assumed to follow independent Poisson distributions with individual rates θ_i, $i = 1, \ldots, 10$, and each pump is observed for a different exposure period E_i, $i = 1, \ldots, 10$. Table 6.6 contains the number of failures and exposure times for each system. An EB analysis for this data set is given by Gaver and O'Muircheartaigh (1987), who employ a log-normal prior distribution for the Poisson rates. The same prior structure is also considered by Carlin and Gelfand (1991), who analyse the data using the Gibbs sampler and rejection sampling to sample from the nonstandard conditional distribution of the Poisson means. Tierney (1994) provides an analysis based on various MCMC methods, while Gelfand and Smith (1990) and George, Makov and Smith (1993) use the Gibbs sampler assuming a conjugate gamma prior distribution.

We used both the approximate Gibbs sampling and the Metropolis-Hastings within Gibbs methods, of Sections 5.4 and 5.5 to analyse the pump failure data. Estimates for the posterior mean and standard deviation of all model parameters are reported in Table 6.7, together with the 2.5% and 97.5% estimated percentiles of the corresponding posterior distributions. In all the examples of this chapter, three independent MCMC chains were combined to ensure convergence of the algorithms.

To compare our results, we also give in Table 6.8 some estimates as pre-

Table 6.7: *Hierarchical Bayes estimates for the pump failure example parameters, using the Metropolis-Hastings within Gibbs and the approximate Gibbs methods.*

		MH within Gibbs				*Gibbs sampling*			
	obs.	*Posterior estimates*				*Posterior estimates*			
par.	*rate*	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1	0.05	0.06	0.02	0.02	0.12	0.06	0.02	0.02	0.12
θ_2	0.06	0.10	0.07	0.01	0.28	0.10	0.07	0.01	0.28
θ_3	0.08	0.09	0.04	0.03	0.17	0.09	0.04	0.03	0.17
θ_4	0.11	0.11	0.03	0.06	0.18	0.12	0.03	0.06	0.18
θ_5	0.57	0.54	0.30	0.13	1.29	0.51	0.29	0.12	1.23
θ_6	0.60	0.60	0.14	0.36	0.89	0.59	0.14	0.36	0.89
θ_7	0.95	0.77	0.74	0.05	2.78	0.67	0.67	0.04	2.53
θ_8	0.95	0.77	0.74	0.05	2.81	0.68	0.68	0.04	2.52
θ_9	1.91	1.63	0.86	0.42	3.71	1.49	0.82	0.35	3.46
θ_{10}	2.10	2.03	0.44	1.27	2.97	2.06	0.44	1.30	3.02
μ		−1.20	0.64	−2.47	0.07	−1.45	0.60	−2.64	−0.21
σ^2		3.63	3.05	0.96	11.20	2.78	2.59	0.63	9.05

Table 6.8: *Posterior means for the pump failure rates, as reported by: Gaver and O'Muircheartaigh (1987), (G&O); Tierney (1994); and George, Makov and Smith (1993), (G,M&S).*

par.	*G&O*	*Tierney*	*G,M&S*
θ_1	0.06	0.07	0.06
θ_2	—	—	0.11
θ_3	—	—	0.09
θ_4	—	—	0.12
θ_5	0.45	0.46	0.60
θ_6	0.57	—	0.61
θ_7	—	—	0.88
θ_8	—	—	0.89
θ_9	—	—	1.56
θ_{10}	1.95	1.91	1.98

sented in earlier analyses of the same data set in the literature. As mentioned earlier, the Gaver and O'Muircheartaigh (1987) estimates are derived using a log-normal prior distribution and following an EB procedure. The results from Tierney (1994) correspond to a log-t prior assumption with 5 degrees of freedom. The parameters of the log-t distribution were set to follow informative hyperprior distributions to match the EB analysis of Gaver and O'Muircheartaigh. Finally, George, Makov and Smith (1993), assume a conjugate Poisson/gamma structure, employing relatively vague priors for the hyperparameters of the gamma distribution. We notice that our estimates are somewhat different from those of previous analyses, reflecting the largely vague uniform distributions that we have assumed at the lower stage of the hierarchical model for the hyperparameters μ and σ^2.

6.2.3 Example: Air conditioning failure data

We also consider a data set concerning the number of air conditioning equipment failures in 13 Boeing 720 aircraft. The observed counts, together with the exposure periods of each aircraft, originally appear in Proschan (1963), and are displayed in Table 6.9. The order in the aircraft number is the same as in Proschan (1963) and corresponds to an ascending order for the observed failure rates.

Table 6.9: *Air conditioning failure data (Proschan, 1963). The exposure times are given in thousands of hours. The order in the aircraft number corresponds to an ascending order for the observed failure rates.*

Aircraft i	Failures y_i	Times E_i
11	2	0.623
9	9	1.800
5	14	1.832
4	15	1.819
12	12	1.297
10	6	0.639
2	23	2.201
3	29	2.422
1	6	0.493
13	16	1.312
7	27	2.074
8	24	1.539
6	30	1.788

Using the modelling in (6.2) and considering both the Metropolis-Hastings within Gibbs and the approximate Gibbs approaches, we obtain the results reported in Table 6.10. The approximate Gibbs method works well, providing estimates that are close to those derived with the hybrid technique. The estimates of the posterior mean of the failure rates are smoothed towards a central value, and therefore lie in a narrower range than the observed rates, which represent the unbiased maximum likelihood estimates. We also notice that the 95% equal tailed intervals reveal a positive skewness for the posterior distribution of the failure rates.

The data have also been analysed by Gaver and O'Muircheartaigh (1987). These authors, considering a log-normal prior on the Poisson rates for their EB approach as in the pump data example, report the posterior means for the rates θ_{11}, θ_2 and θ_6 as 8.67, 10.38 and 13.60 respectively, and the maximum likelihood estimates for the hyperparameters μ and σ^2 as 2.34 and 0.053. These estimates demonstrate that the EB approach results in more radical shrinkages when compared to the hierarchical Bayes analysis, due to the fact that the former method does not adequately account for the uncertainty in the hyperparameters.

Table 6.10: *Hierarchical Bayes estimates for the air conditioning failure example parameters, using the Metropolis-Hastings within Gibbs and the approximate Gibbs methods.*

		MH within Gibbs				*Gibbs sampling*			
	obs.	*Posterior estimates*				*Posterior estimates*			
par.	*rate*	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_{11}	3.21	7.73	2.43	3.34	12.65	7.66	2.43	3.19	12.64
θ_9	5.00	7.14	1.81	3.95	10.94	7.10	1.81	3.83	10.83
θ_5	7.64	8.66	1.84	5.28	12.44	8.64	1.82	5.34	12.40
θ_4	8.25	9.05	1.82	5.77	12.93	9.02	1.84	5.66	12.88
θ_{12}	9.25	9.77	2.12	5.96	14.26	9.69	2.15	5.90	14.37
θ_{10}	9.39	10.03	2.63	5.51	15.94	10.03	2.68	5.36	16.0
θ_2	10.45	10.42	1.82	7.14	14.28	10.43	1.82	7.16	14.33
θ_3	11.97	11.50	1.91	8.18	15.59	11.51	1.90	8.17	15.56
θ_1	12.17	11.13	3.08	6.09	18.34	11.14	3.13	6.06	18.54
θ_{13}	12.20	11.44	2.36	7.44	16.69	11.44	2.38	7.34	16.72
θ_7	13.02	12.16	2.08	8.54	16.64	12.18	2.12	8.55	16.81
θ_8	15.59	13.75	2.67	9.32	19.65	13.68	2.63	9.34	19.52
θ_6	16.78	14.67	2.71	10.18	20.59	14.69	2.70	10.16	20.75
μ		2.31	0.13	2.03	2.56	2.31	0.13	2.02	2.55
σ^2		0.15	0.13	0.01	0.49	0.15	0.14	0.02	0.52

6.2.4 Random effects model

We now consider the situation where the means of the Poisson distributions depend on p explanatory variables. If no different exposure times are involved, we assume that the Poisson means are given as

$$\begin{aligned}\theta_i &= \exp(b_0 x_{0i} + b_1 x_{1i} + \ldots + b_{p-1} x_{p-1,i} + \varepsilon_i) \\ &= \exp(\mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \varepsilon_i), \qquad i = 1, \ldots, m,\end{aligned} \tag{6.3}$$

where $\mathbf{x}_i^{\mathrm{T}} = (x_{0i}, x_{1i}, \ldots, x_{p-1,i})$, for $i = 1, \ldots, m$, are known design vectors, $\mathbf{b} = (b_0, b_1, \ldots, b_{p-1})^{\mathrm{T}}$ is a vector of regression coefficients and ε_i, $i = 1, \ldots, m$, are random error terms with means zero and constant unknown variance σ^2. Then we can define a generalised linear model with random effects, expressed as

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \gamma_i = \log(\theta_i) &= \mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \varepsilon_i \\ \varepsilon_i &\overset{iid}{\sim} N(0, \sigma^2) \\ \pi(\mathbf{b}) &\propto 1 \\ \pi(\sigma^2) &\propto 1 \end{aligned} \tag{6.4}$$

where $i = 1, \ldots, m$. Here, we have assumed that the coefficients $b_0, b_1, \ldots, b_{p-1}$, and the variance σ^2 are independently distributed according to vague uniform prior distributions over their range. The full Bayesian analysis of such models has been mainly tackled with the use of the Gibbs sampler combined with various rejection sampling techniques, following the work of Zeger and Karim (1991). Nevertheless, we can rearrange the formulation in model (6.4), in such a way that we can exploit our approximation to the full conditional distribution of the logarithm of the Poisson means, derived in Chapter 5. The model may be presented as

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \gamma_i &= \log(\theta_i) \\ \gamma_i|\lambda_i, \sigma^2 &\overset{ind}{\sim} N(\lambda_i, \sigma^2) \\ \lambda_i &= \mathbf{x}_i^{\mathrm{T}}\mathbf{b} \\ \pi(\mathbf{b}) &\propto 1 \\ \pi(\sigma^2) &\propto 1 \end{aligned} \tag{6.5}$$

for $i = 1, \ldots, m$. Then, the above model gives the following joint posterior distribution for $\boldsymbol{\gamma}$, $\mathbf{b}$ and σ^2

$$p(\boldsymbol{\gamma}, \mathbf{b}, \sigma^2|\mathbf{y}) \propto (\sigma^2)^{-\frac{1}{2}m} \exp\left[\sum_{i=1}^{m}\left\{\gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\sigma^{-2}(\gamma_i - \lambda_i)^2\right\}\right], \tag{6.6}$$

which is the same as the one for model (4.1a) with μ replaced by $\lambda_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{b}$. This implies that the full conditional distributions for σ^2 and γ_i, $i = 1, \ldots, m$, are the of the same form as those given in (5.44) and (5.45) respectively, i.e.

$$\frac{\sum_{i=1}^{m}(\gamma_i - \lambda_i)^2}{\sigma^2} \Big| \boldsymbol{\gamma}, \mathbf{b}, \mathbf{y} \sim \chi^2_{m-2}. \tag{6.7}$$

and

$$p(\gamma_i|\mathbf{b}, \sigma^2, \mathbf{y}) \propto \exp\left\{\gamma_i y_i - e^{\gamma_i} - \frac{1}{2}\,\sigma^{-2}(\gamma_i - \lambda_i)^2\right\}, \qquad i = 1 \ldots, m, \tag{6.8}$$

where $\lambda_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{b}$ for $i = 1, \ldots, m$.

We will now derive the full conditional distribution for the regression coefficients b_k, $k = 0, \ldots, p-1$. We use the notation $\mathbf{b}_{-k}$ to indicate the vector $\mathbf{b}$ with its kth component omitted, that is

$$\mathbf{b}_{-k} = (b_0, \ldots, b_{k-1}, b_{k+1}, \ldots, b_{p-1})^{\mathrm{T}}. \tag{6.9}$$

We also denote the corresponding linear component by h_{ki}, i.e.

$$h_{ki} = b_0 x_{0i} + \ldots + b_{k-1}x_{k-1,i} + b_{k+1}x_{k+1,i} + \ldots + b_{p-1}x_{p-1,i}. \tag{6.10}$$

Then, the full conditional distribution of b_k given $\boldsymbol{\gamma}$, $\mathbf{b}_{-k}$, σ^2 and $\mathbf{y}$ is given from (6.6) as

$$p(b_k|\boldsymbol{\gamma}, \mathbf{b}_{-k}, \sigma^2, \mathbf{y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(\gamma_i - \mathbf{x}_i^{\mathrm{T}}\mathbf{b})^2\right\},$$

and using the decomposition $\mathbf{x}_i^{\mathrm{T}}\mathbf{b} = h_{ki} + b_k x_{ki}$ and ignoring terms that do not involve b_k, we can write

$$\begin{aligned} p(b_k|\boldsymbol{\gamma}, \mathbf{b}_{-k}, \sigma^2, \mathbf{y}) &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m}\{(\gamma_i - h_{ki}) - b_k x_{ki}\}^2\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{m} b_k^2 x_{ki}^2 - 2b_k\sum_{i=1}^{m}(\gamma_i - h_{ki})x_{ki}\right\}\right] \\ &= \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m} x_{ki}^2\left\{b_k^2 - 2b_k\frac{\sum_{i=1}^{m}(\gamma_i - h_{ki})x_{ki}}{\sum_{i=1}^{m} x_{ki}^2}\right\}\right]. \end{aligned}$$

Completing the square and ignoring again terms not involving b_k, we obtain

$$p(b_k|\boldsymbol{\gamma}, \mathbf{b}_{-k}, \sigma^2, \mathbf{y}) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{m} x_{ki}^2\left\{b_k - \frac{\sum_{i=1}^{m}(\gamma_i - h_{ki})x_{ki}}{\sum_{i=1}^{m} x_{ki}^2}\right\}^2\right]. \tag{6.11}$$

Expression (6.11) indicates that conditionally on $\boldsymbol{\gamma}$, $\mathbf{b}_{-k}$, σ^2 and $\mathbf{y}$, b_k follows a normal distribution, and we can write

$$b_k|\boldsymbol{\gamma}, \mathbf{b}_{-k}, \sigma^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^{m}(\gamma_i - h_{ki})x_{ki}}{\sum_{i=1}^{m} x_{ki}^2}, \frac{\sigma^2}{\sum_{i=1}^{m} x_{ki}^2}\right), \tag{6.12}$$

for $k = 0, \ldots, p-1$, with h_{ki} given in (6.10).

We notice here that if the whole vector $\mathbf{b}$ is considered as a block, working in matrix formulation in a manner similar to that leading to (6.11), we derive a multivariate normal conditional distribution of the form

$$\mathbf{b}|\boldsymbol{\gamma}, \sigma^2, \mathbf{y} \sim N_p\left(\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{m}\mathbf{x}_i\gamma_i,\ \sigma^2\left(\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\right)^{-1}\right).$$

Therefore, we can employ the above, or (6.12), together with the full conditionals in (6.7) and (6.8) to proceed with the Gibbs sampling and Metropolis-Hastings within Gibbs methods presented in Chapter 5.

6.2.5 Random effects model for the analysis of event rates

The generalised linear model of the preceding subsection can be extended by allowing the observations to involve exposure times, or denominators E_i, so that the Poisson means are given by $\mu_i = \theta_i E_i$, $i = 1, \ldots, m$, as in Subsection 6.2.1. Then the log-linear form in (6.3) corresponds to the Poisson rates θ_i, $i = 1, \ldots, m$. Thus, when the conditional sampling distribution of Y_i given θ_i in model (6.5) is substituted by a Poisson$(\theta_i E_i)$ distribution, the full conditional density $p(\gamma_i|\mathbf{b}, \sigma^2, \mathbf{y})$ in (6.8) must be adjusted, and the approximation methods cannot be applied. However, we may employ the equivalent model formulation

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\mu_i) \\ \gamma_i &= \log(\mu_i) \\ \gamma_i|\lambda_i^*, \sigma^2 &\overset{ind}{\sim} N(\lambda_i^*, \sigma^2) \\ \lambda_i^* &= \log E_i + \mathbf{x}_i^{\mathrm{T}}\mathbf{b} \\ \pi(\mathbf{b}) &\propto 1 \\ \pi(\sigma^2) &\propto 1 \end{aligned} \tag{6.13}$$

for $i = 1, \ldots, m$, which allows for the denominators E_i, $i = 1, \ldots, m$, avoiding any changes in the formulation of the model given in (6.5). It follows that we can proceed as for the usual random effects model presented in Subsection 6.2.4, employing the full conditional distributions in (6.12), (6.7) and (6.8) for the MCMC methods. To do so, we only need to replace h_{ki} and λ_i in these expressions by

$$h_{ki}^* = \log E_i + h_{ki}$$

Table 6.11: *Heart transplant data (Christiansen and Morris, 1997).*

Centre i	No. of patients E_i	No. of deaths y_i	Severity x_{1i}
1	18	4	0.251
2	24	1	−0.021
3	28	3	0.208
4	37	1	0.019
5	42	2	−0.169
6	48	1	0.164
7	56	2	0.296
8	60	3	0.199
9	68	6	0.209
10	73	0	0.093
11	75	8	0.002
12	79	3	0.064
13	91	3	0.105
14	99	9	0.073
15	104	7	0.209

and λ_i^* from (6.13) respectively. The Poisson rates θ_i can be obtained by dividing the derived Poisson means μ_i by the denominators E_i, for $i = 1, \ldots, m$.

6.2.6 Example: Heart transplant data

We consider a data set analysed in Christiansen and Morris (1997). It concerns the mortality rates in $m = 15$ U.S. heart transplant centres in a period between October 1989 and December 1989, taken as a sample from 131 centres considered in Christiansen and Morris (1996). The number of deaths y_i, within 30 days of heart transplant surgery out of E_i, $i = 1, \ldots, 15$, treated patients at each centre were recorded. Also available is a covariate x_{i1}, $i = 1, \ldots, 15$, which is referred to as the severity index, and is based on seven risk variables for each patient treated at centre i, thus reflecting the severity of all the cases treated at each centre. The data are presented in Table 6.11.

We can obtain a hierarchical Bayes analysis for these data using model (6.13), where $\mathbf{x}_i^{\mathrm{T}} = (x_{0i}, x_{1i})$, with $x_{0i} = 1$ for $i = 1, \ldots, 15$, and x_{1i} representing the severity index. The denominators E_i, $i = 1, \ldots, 15$, can be viewed as exposures leading to observed Poisson rates $\frac{y_i}{E_i}$, with the Poisson approximation to the binomial distribution justified from the fact that the observed proportion of deaths was less than 0.08 in the original data set.

The results from the hierarchical Bayes analysis are shown in Table 6.12,

Table 6.12: *Hierarchical Bayes posterior estimates for the heart transplant mortality example parameters. The results from both the Metropolis-Hastings within Gibbs and the approximate Gibbs methods are given.*

		MH within Gibbs				*Gibbs sampling*			
	obs.	*Posterior estimates*				*Posterior estimates*			
par.	*rate*	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1	0.222	0.145	0.082	0.042	0.355	0.146	0.083	0.041	0.355
θ_2	0.042	0.046	0.030	0.008	0.122	0.046	0.030	0.007	0.125
θ_3	0.107	0.087	0.045	0.026	0.199	0.088	0.046	0.026	0.201
θ_4	0.027	0.040	0.023	0.008	0.097	0.040	0.024	0.007	0.098
θ_5	0.480	0.044	0.026	0.010	0.110	0.044	0.027	0.009	0.111
θ_6	0.021	0.038	0.022	0.007	0.090	0.038	0.022	0.006	0.092
θ_7	0.036	0.049	0.026	0.012	0.109	0.049	0.026	0.011	0.108
θ_8	0.050	0.054	0.025	0.017	0.112	0.054	0.025	0.016	0.115
θ_9	0.088	0.080	0.030	0.034	0.151	0.080	0.030	0.033	0.151
θ_{10}	0.000	0.023	0.016	0.002	0.061	0.023	0.016	0.002	0.060
θ_{11}	0.107	0.089	0.032	0.040	0.164	0.089	0.033	0.039	0.165
θ_{12}	0.038	0.043	0.019	0.014	0.087	0.043	0.019	0.013	0.087
θ_{13}	0.033	0.040	0.018	0.013	0.081	0.039	0.018	0.012	0.080
θ_{14}	0.091	0.081	0.026	0.040	0.141	0.081	0.027	0.039	0.142
θ_{15}	0.067	0.066	0.023	0.030	0.118	0.066	0.023	0.029	0.118
b_0		−3.18	0.40	−4.04	−2.47	−3.18	0.40	−4.06	−2.47
b_1		1.36	2.32	−3.06	6.11	1.39	2.33	−3.07	6.13
σ^2		0.86	1.07	0.07	3.30	0.91	1.22	0.06	3.53

Table 6.13: *Posterior estimates for the heart transplant mortality example parameters as given in Christiansen and Morris (1997).*

par.	*mean*	*sd*	*par.*	*mean*	*sd*
θ_1	0.117	0.057	θ_9	0.079	0.028
θ_2	0.049	0.028	θ_{10}	0.024	0.016
θ_3	0.083	0.038	θ_{11}	0.083	0.029
θ_4	0.043	0.023	θ_{12}	0.045	0.019
θ_5	0.047	0.027	θ_{13}	0.042	0.018
θ_6	0.042	0.022	θ_{14}	0.079	0.024
θ_7	0.051	0.026	θ_{15}	0.067	0.022
θ_8	0.057	0.024			

where we report posterior estimates of the model parameters, derived with the Gibbs sampling and Metropolis-Hastings within Gibbs methods. For comparison reasons we also present the posterior mean and standard deviation estimates of the Poisson rates from the Christiansen and Morris (1997) analysis, in Table 6.13. These authors consider a conjugate Poisson/gamma hierarchical model, and they rely on various approximations to obtain a Bayesian analysis. At the second prior stage they assume vague flat distributions on the regression coefficients and a uniform $U(0, 1)$ hyperprior distribution on the shrinkage coefficient, in a manner similar to the modelling structure given in (4.2).

Our hierarchical Bayes estimates of the mortality rates are smoothed towards the centre of the observed rates, and the estimate of the regression coefficient b_1 suggests that large values of a centre's severity index will increase the expected mortality rate at that centre. The comparison between the results from the Poisson/log-normal and the Poisson/gamma model reveals that both structures produced similar posterior estimates. However, the analysis of the former model resulted in slightly smaller shrinking and larger posterior variation for the mortality rates, reflecting again the largely vague hyperprior distributions in our analysis.

6.2.7 Example: Lip cancer in Scotland data

Clayton and Kaldor (1987) analyse a data set which involves lip cancer incidents in $m = 56$ counties in Scotland during the years between 1975 and 1980. The observed and expected lip cancer cases are reported, denoted as y_i and E_i, $i = 1, \ldots, 56$, respectively, with the expected numbers calculated by the authors based on the age distribution in each county, and treated as known constants in

the analysis. A measure of the area specific relative risk is given by the ratio $\frac{y_i}{E_i}$, the so-called standardised mortality ratio (SMR). Under the assumption that the random variables Y_i follow conditionally independent Poisson distributions, given the unobserved rates θ_i, with respective means $\theta_i E_i$, for $i = 1, \ldots, 56$, the observed SMR give the ML estimates for θ_i, $i = 1, \ldots, m$.

Clayton and Kaldor considered various assumptions for the prior distribution of the area specific relative risks, that is the Poisson rates θ_i, $i = 1, \ldots, m$, including a log-normal prior to accommodate the assumed spatial correlation, and adopted an EB approach for the estimation problem. Breslow and Clayton (1993) also consider the same data set, including a covariate representing the percentage of population employed in agriculture, fishing or forestry in each county, and followed a quasi-likelihood procedure. It is believed that the covariate can explain a large part of the spatial aggregation. The authors also consider a model structure permitting intrinsic spatial correlation, omitting the explanatory variable in order to make their analysis comparable to that by Clayton and Kaldor. We will examine the model including the covariate, but assuming uncorrelated random effects. The observed and expected lip cancer cases, together with the covariate values x_{1i} for counties $i = 1, \ldots, 56$, are given in Table 6.14.

Breslow and Clayton assume a log-linear form for the Poisson means μ_i, $i = 1, \ldots, 56$, given as

$$\log(\mu_i) = \log E_i + b_0 + b_1 \frac{x_i}{10} + \varepsilon_i, \qquad i = 1, \ldots, 56,$$

where the random effects ε_i are independently distributed according to a normal distribution with mean zero and variance σ^2. Hence, assuming vague prior distributions for the regression coefficients and σ^2, we can express our model as

$$\begin{aligned}
Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\mu_i) \\
\gamma_i &= \log(\mu_i) \\
\gamma_i|\lambda_i^*, \sigma^2 &\overset{ind}{\sim} N(\lambda_i^*, \sigma^2) \\
\lambda_i^* &= \log E_i + b_0 + b_1 \frac{x_i}{10} \\
\pi(\mathbf{b}) &\propto 1 \\
\pi(\sigma^2) &\propto 1
\end{aligned}$$

for $i = 1, \ldots, 56$, which is model (6.13). Therefore, we can proceed as described in Subsection 6.2.5 to obtain hierarchical Bayes estimates for the relative risks $\theta_i = \frac{\mu_i}{E_i}$, $i = 1, \ldots, 56$. Table 6.15 contains posterior estimates for the model parameters, derived with the two proposed MCMC methods. We note that the observed relative risks and all corresponding estimates are multiplied by a factor of 100, to conform with a convention traditionally adopted for the SMR measure.

Table 6.14: *Lip cancer in Scotland data (Breslow and Clayton, 1993).*

County i	Obs. y_i	Exp. E_i	Cov. x_i	County i	Obs. y_i	Exp. E_i	Cov. x_i
1	9	1.4	16	29	16	14.4	10
2	39	8.7	16	30	11	10.2	10
3	11	3.0	10	31	5	4.8	7
4	9	2.5	24	32	3	2.9	24
5	15	4.3	10	33	7	7.0	10
6	8	2.4	24	34	8	8.5	7
7	26	8.1	10	35	11	12.3	7
8	7	2.3	7	36	9	10.1	0
9	6	2.0	7	37	11	12.7	10
10	20	6.6	16	38	8	9.4	1
11	13	4.4	7	39	6	7.2	16
12	5	1.8	16	40	4	5.3	0
13	3	1.1	10	41	10	18.8	1
14	8	3.3	24	42	8	15.8	16
15	17	7.8	7	43	2	4.3	16
16	9	4.6	16	44	6	14.6	0
17	2	1.1	10	45	19	50.7	1
18	7	4.2	7	46	3	8.2	7
19	9	5.5	7	47	2	5.6	1
20	7	4.4	10	48	3	9.3	1
21	16	10.5	7	49	28	88.7	0
22	31	22.7	16	50	6	19.6	1
23	11	8.8	10	51	1	3.4	1
24	7	5.6	7	52	1	3.6	0
25	19	15.5	1	53	1	5.7	1
26	15	12.5	1	54	1	7.0	1
27	7	6.0	7	55	0	4.2	16
28	10	9.0	7	56	0	1.8	10

Table 6.15: *Hierarchical Bayes posterior estimates for the lip cancer in Scotland example parameters. The results from both the Metropolis-Hastings within Gibbs and the approximate Gibbs methods are given.*

		MH within Gibbs				*Gibbs sampling*			
	obs.	*Posterior estimates*				*Posterior estimates*			
par.	*rate*	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_1	652.2	484.9	166.6	224.2	873.2	482.2	165.5	227.1	869.1
θ_2	450.3	424.7	67.7	301.7	566.9	425.3	68.4	302.8	569.2
θ_3	361.8	298.6	90.5	150.3	504.0	298.7	89.9	151.5	500.0
θ_4	355.7	353.2	106.8	176.9	597.0	353.3	107.7	177.4	590.1
θ_5	352.1	300.2	78.2	172.0	474.5	298.8	78.6	169.2	469.9
θ_6	333.3	332.5	105.0	163.2	568.3	332.0	105.6	161.5	573.9
θ_7	320.6	294.1	58.0	191.9	418.6	295.1	57.6	194.3	421.3
θ_8	304.3	223.6	84.2	95.7	419.8	226.3	85.1	94.8	425.1
θ_9	303.0	214.5	85.2	87.7	417.9	215.9	86.5	87.1	412.8
θ_{10}	301.7	288.1	62.8	180.6	424.4	287.7	63.2	178.4	424.5
θ_{11}	295.5	247.4	69.2	133.5	404.2	246.0	68.0	133.4	402.2
θ_{12}	279.3	246.5	96.9	101.5	480.6	249.3	98.1	100.1	473.2
θ_{13}	277.8	194.8	94.4	63.9	421.5	195.4	96.2	62.0	427.1
θ_{14}	241.7	260.0	81.1	129.0	439.9	259.2	80.5	127.8	442.0
θ_{15}	216.8	197.9	47.1	116.7	300.7	197.1	47.2	116.9	299.0
θ_{16}	197.8	194.0	58.3	99.0	325.3	193.6	57.9	99.4	320.7
θ_{17}	186.9	154.5	80.8	47.5	359.0	154.2	79.1	44.4	351.1
θ_{18}	167.5	146.9	50.9	66.9	262.6	146.9	50.9	66.6	262.4
θ_{19}	162.7	148.7	46.2	73.7	253.3	147.3	46.4	74.6	253.8
θ_{20}	157.7	150.8	51.0	69.2	268.0	150.3	50.2	68.3	265.1
θ_{21}	153.0	144.3	34.7	84.9	220.8	144.6	34.4	84.0	220.0
θ_{22}	136.7	139.5	24.3	96.6	191.2	139.5	24.1	97.1	190.1
θ_{23}	125.4	124.8	34.1	67.6	201.0	124.6	34.1	67.3	201.5
θ_{24}	124.6	119.4	39.5	55.9	208.7	118.7	39.2	54.5	209.2
θ_{25}	122.8	114.0	25.6	69.7	170.3	114.1	25.5	70.4	169.4
θ_{26}	120.1	110.7	27.6	64.1	171.4	110.5	27.7	63.4	172.1
θ_{27}	115.9	112.5	37.2	52.7	198.4	112.5	37.6	52.7	199.8
θ_{28}	111.6	109.2	31.2	57.6	179.9	109.1	31.1	57.0	177.2

Table 6.15: *(Continued)*

		MH within Gibbs				*Gibbs sampling*			
	obs.	*Posterior estimates*				*Posterior estimates*			
par.	*rate*	*mean*	*sd*	2.5%	97.5%	*mean*	*sd*	2.5%	97.5%
θ_{29}	111.3	112.8	26.3	67.6	170.7	112.5	26.3	66.5	168.1
θ_{30}	107.8	110.7	30.1	60.4	177.8	110.8	30.2	60.2	177.0
θ_{31}	105.3	104.1	38.7	44.2	193.5	104.0	39.1	43.7	191.9
θ_{32}	104.2	163.0	64.3	63.9	311.6	162.6	64.0	63.5	310.7
θ_{33}	99.6	106.2	34.4	50.9	183.5	106.2	34.2	51.8	182.4
θ_{34}	93.8	95.6	29.6	48.0	161.9	95.8	29.9	47.4	162.6
θ_{35}	89.3	91.4	25.0	49.3	146.8	91.5	25.1	50.2	146.7
θ_{36}	89.1	83.0	25.6	41.4	140.7	82.8	24.8	41.7	140.4
θ_{37}	86.8	92.6	24.9	50.9	148.1	92.2	24.8	50.5	146.2
θ_{38}	85.6	81.2	26.0	39.1	139.6	80.3	25.5	39.2	139.6
θ_{39}	83.3	103.6	33.6	49.5	179.7	103.6	33.6	48.7	178.6
θ_{40}	75.9	71.5	29.0	28.3	140.3	71.8	29.3	27.3	141.2
θ_{41}	53.3	55.6	15.6	29.7	90.6	55.7	15.4	29.7	89.8
θ_{42}	50.7	66.4	18.8	34.9	107.4	66.5	19.1	34.9	109.2
θ_{43}	46.3	89.7	37.8	32.7	178.8	89.7	37.3	32.8	177.8
θ_{44}	41.0	46.2	15.4	21.4	80.9	46.4	15.6	21.1	80.7
θ_{45}	37.5	39.9	8.5	25.2	58.3	40.0	8.5	24.8	57.7
θ_{46}	36.6	55.2	21.5	21.8	104.8	55.4	21.3	21.9	105.6
θ_{47}	35.8	50.3	22.5	17.4	103.1	50.3	22.8	17.0	104 .0
θ_{48}	32.1	44.2	17.6	17.5	85.5	44.3	17.6	17.0	84.9
θ_{49}	31.6	33.2	5.9	22.8	45.8	33.2	5.9	22.7	45.8
θ_{50}	30.6	37.6	12.3	18.0	65.1	37.9	12.3	17.8	65.4
θ_{51}	29.1	52.1	26.5	15.6	118.0	52.5	26.3	14.8	119.1
θ_{52}	27.6	48.8	24.9	14.6	111.4	49.1	24.9	14.0	110.3
θ_{53}	17.4	41.0	19.5	12.9	87.7	41.0	19.4	12.6	87.1
θ_{54}	14.2	36.6	17.2	11.9	78.4	36.9	17.3	11.6	78.2
θ_{55}	0.0	64.2	31.1	19.0	138.3	63.7	30.7	18.2	137.9
θ_{56}	0.0	76.2	41.5	20.2	179.8	76.3	41.3	19.0	176.9
β_0		−0.51	0.15	−0.80	−0.22	−0.50	0.15	−0.81	−0.22
β_1		0.69	0.13	0.42	0.95	0.68	0.13	0.42	0.95
σ^2		0.45	0.14	0.24	0.79	0.45	0.14	0.24	0.79

Table 6.15 shows that the hierarchical Bayes estimator draws the observed SMRs towards a central value, correcting the extreme relative risks of counties based on small observed counts. The estimates of the regression coefficients b_0 and b_1 are similar to those given by Breslow and Clayton, that is -0.44 and 0.68 respectively. Our estimate for σ^2 suggests that their method underestimates the variance component, giving a value of 0.36.

6.2.8 Further extensions

The situations that we have modelled can be further extended in many various ways, to represent different and complicated aspects of real life problems. In the present subsection we will illustrate the potential of the modelling presented in this thesis, by briefly mentioning two more directions towards which similar models can be extended, leaving the analysis details as possible work for future research.

Poisson model for genetic traits

Tempelman and Gianola (1993) present a model for genetic reproductive traits in dairy science. The observed counts $y_1, y_2, \ldots, y_m$, which for example can represent litter size, are assumed to follow independent Poisson distributions, given their respective means $\theta_1, \theta_2, \ldots, \theta_m$. The latter are assumed to depend on fixed and random effects through a log-linear component of the form

$$\gamma_i = \log(\theta_i) = \mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\lambda}, \qquad i = 1, \ldots, m,$$

where $\mathbf{x}_i$ and $\mathbf{u}_i$ are specified $p \times 1$ and $q \times 1$ design vectors, $\mathbf{b}$ is a $p \times 1$ vector of fixed effects, and $\boldsymbol{\lambda}$ is a $q \times 1$ vector of random effects. Furthermore, we assume that $\boldsymbol{\lambda}$ follows a q-dimensional multivariate normal distribution with zero mean vector and covariance matrix $\sigma^2\mathbf{A}$. At the lower stage of the Bayesian hierarchical model, once again we assume vague uniform prior distributions for $b_0, b_1, \ldots, b_{p-1}$, and σ^2 over their range. Thus, the model can be given as

$$\begin{aligned} Y_i|\theta_i &\overset{ind}{\sim} \text{Poisson}(\theta_i) \\ \gamma_i = \log(\theta_i) &= \mathbf{x}_i^{\mathrm{T}}\mathbf{b} + \mathbf{u}_i^{\mathrm{T}}\boldsymbol{\lambda} \\ \boldsymbol{\lambda} &\sim N_q(\mathbf{0}, \sigma^2\mathbf{A}) \\ \pi(\mathbf{b}) &\propto 1 \\ \pi(\sigma^2) &\propto 1 \end{aligned} \tag{6.14}$$

for $i = 1, \ldots, m$. We notice that $\mathbf{A}$ is an additive genetic relationship matrix which can be specified by the geneticists, and σ^2, the so-called additive genetic

variance, is the parameter of primary interest. Tempelman and Gianola (1993) propose Laplacian approximations that lead to approximate inferences for the variance component σ^2. The analysis of model (6.14) in a similar manner to that of model (6.5) can be the subject of further investigation, together with other approaches combining analytical approximations and Monte Carlo integration techniques for the exact inference.

Spatial correlation model

In the random effect log-linear model (6.4) the random effects ε_i, $i = 1, \ldots, m$, were assumed to be independent. However, in situations where, for instance, counts are observed in geographical regions, the existence of some spatial dependence between neighbouring areas may be expected, thus leading to the assumption of correlated random effects. The spatial correlation can be conveniently accommodated through an assumed multivariate first stage prior distribution, specifying a suitable form for the mean vector and covariance matrix with the conditional individual means and/or variances depending on the means and variances of geographically neighbouring components.

In a general setting for a model where correlation arises from geographical proximity, we can assume that each of the random effects ε_i, $i = 1, \ldots, m$, in model (6.4) conditionally follows a normal distribution depending on the remaining ε_j, $j \neq i$. If we use $\boldsymbol{\varepsilon}_{-i}$ to denote the vector consisting of all ε_j, $j = 1, \ldots, m$, except ε_i, and suppressing the dependence on any other model parameters, the mean of ε_i can be given as

$$\mathrm{E}(\varepsilon_i | \boldsymbol{\varepsilon}_{-i}) = \mu(\boldsymbol{\varepsilon}_{-i}), \qquad i = 1, \ldots, m,$$

where $\mu(\boldsymbol{\varepsilon}_{-i})$ denotes a function of the components of $\boldsymbol{\varepsilon}_{-i}$. Also the covariance between ε_i and ε_j can be written as

$$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \, h\{d(i,j)\}, \qquad i, j = 1, \ldots, m,$$

where $h(\cdot)$ is a function of a distance measure $d(i,j)$ between i and j.

As mentioned in Subsection 6.2.7, Breslow and Clayton (1993) present a random effect log-linear model for the lip cancer in Scotland data, considering spatially correlated random effects. They assume that the underlying dependence among the area specific relative risks can be expressed through the neighbouring relationship among individual areas, introducing an intrinsic normal prior for the random effects. The model can be expressed as a modification to the model

described in Subsection 6.2.7, that is

$$Y_i|\theta_i \overset{ind}{\sim} \text{Poisson}(\mu_i)$$
$$\gamma_i = \log(\mu_i) = \log E_i + b_0 + b_1 \frac{x_i}{10} + \varepsilon_i$$
$$\varepsilon_i|\boldsymbol{\varepsilon}_{-i} \sim N(\bar{\varepsilon}_i, \frac{\sigma^2}{n_i})$$
$$\pi(\mathbf{b}) \propto 1$$
$$\pi(\sigma^2) \propto 1$$

for $i = 1, \ldots, m$, where n_i is the number of the neighbours of i and $\bar{\varepsilon}_i = \frac{1}{n_i} \sum_{j \sim i} \varepsilon_j$, with $j \sim i$ denoting adjacent counties. The analysis of the above model, or relevant models allowing for spatial correlation, in a framework similar to that of the present thesis, would also be an interesting subject of further research.

6.3 Summary and conclusions

In the present chapter we have considered the average risk of the hierarchical Bayes estimator. We have also investigated the possibility of applying the estimating methods presented in Chapter 5, to extended models.

The average risk of the hierarchical Bayes estimator was examined in a wide range of the parameter space and for different loss functions. Due to the computationally intensive nature of the frequency simulation study, special attention should be given to the convergence issue of the MCMC methods, which was accelerated employing the EB estimating methods presented in previous chapters to obtain suitable initial algorithm values.

Both model specifications (4.1a) and (4.1b) were considered. With the former we assume vague uniform priors at the lower prior stage, and thus the hierarchical Bayes estimator provides very good frequency properties in almost all the examined parameter regions, including the cases when the variance of the parameters θ_i, $i = 1, \ldots, m$, is relatively large. Under the more informative hyperprior distribution for the variance parameter σ^2 in model (4.1b), the hierarchical Bayes analysis results in an estimator with average risk properties which are comparable to those of the EB estimators. It gives great improvement over the average risk of the unbiased estimator when the true parameters are closely concentrated to each other, but does not perform equally well in the presence of larger variation among them. Finally, the frequency properties of the hierarchical Bayes estimator obtained with the approximate Gibbs sampling method were also examined, with the results being similar as before.

The applicability of our estimating methods to more complicated structures was also explored. The basic Poisson models considered in the preceding chapters

can be extended to represent various situations where extra-Poisson variation occurs. We present possible extensions where events can be considered to occur, for example, over unequal exposure periods, in which case the event rates over a unit interval, rather than the actual Poisson means, are of interest. We also discuss the case where explanatory variables are employed to account for the observed overdispersion. The model is altered to allow the logarithms of the Poisson means to depend on the covariates through a linear component in a regression structure assuming a normally distributed random error term. The resulting random effect generalised linear model may also be taken to include different exposures, as described above.

In all cases, suitable modelling was employed to ensure that the approximations to the full conditional distribution of the Poisson parameters, as developed in Chapter 5, were still effective. Hence, the described hierarchical Bayes estimating methods could be applied, with only a few minor adjustments, as illustrated through various examples. Again, the approximate Gibbs method performed well, providing estimates which were very close to those from the exact method. Finally, scope for further research was given, with the presentation of two further extended cases.

Chapter 7

Conclusions

The aim of this thesis was to develop inference methods for models accounting for extra-Poisson variation, under an empirical and hierarchical Bayesian framework. Another issue addressed, was the assessment of the methods in terms of the average risk of the resulting estimators under various loss functions. Nonconjugate Poisson/log-normal models were mainly considered, and both vague and informative prior structures were assumed at the lower stage of the hierarchical model. Methods for the implementation of simulation-based techniques were suggested, and their extension to more complicated structures was discussed.

The emergence of linear shrinkage rules in the problem of the simultaneous estimation of the means of several conditionally independent Poisson distributions, as motivated by the inadmissibility of the unbiased estimator, was reviewed in Chapter 2 under both the frequentist and the Bayesian philosophy. Various methods were presented, illustrating the issue of constructing estimators that shrink towards predetermined directions and perform well under a specific loss function.

In Chapter 3 we introduced a Poisson/log-normal one stage Bayesian model. The log-normal prior assumption allows for more variation in the Poisson parameters, in relation to the conjugate gamma distribution, and it can also accommodate a correlation structure. Thus, it was preferred to the usual gamma prior specification, at the expense of more complicated computations. For the estimation of the parameters of interest we focused on two methods approximating the posterior mean, conditional on the parameters of the prior distribution. We first described the construction of the best linear predictor, which is a Bayes linear rule that was developed in a such a way that it possesses the minimum average risk among all linear estimators of the same form. Then, an importance sampling method was proposed, based on an appropriate gamma importance density, which resulted in a very accurate approximation to the posterior mean. Since both methods assumed knowledge of the model hyperparameters, an empirical Bayes solution was developed. It was shown by means of data examples that

the empirical Bayes posterior mean shrinks the usual estimates towards a central point, which lies between the sample mean and the minimum observation.

We also examined the frequency properties of the suggested estimators using the criterion of average risk. A simulation study was conducted to assess the average risk under different loss functions, including weighted squared and absolute error loss, summed over the components of the parameter vector and also considering the maximum loss component. Different regions of the parameter space were explored, with the results demonstrating that the presented estimators provide considerable improvement when compared to the usual unbiased estimator and the methods reviewed in Chapter 2, in most of the examined cases.

A full Bayesian hierarchical Poisson structure was presented in Chapter 4. Some analytical approximations for the estimation of the posterior mean for both the Poisson/gamma and the Poisson/log-normal models were given. In the latter formulation, we assumed the flat uniform prior distributions $U(-\infty, \infty)$ and $U(0, \infty)$ for the hyperparameters μ and σ^2 respectively. We relied on a normal approximation to the marginal distribution of the data to facilitate the computations arising from densities of nonclosed form. Further approximations were employed, but fully analytical expressions were not available and the derivation of the final estimates still requires the use of numerical integration. However, the resulting expressions provide a general indication of the nature of the posterior mean, also suggesting that the hierarchical Bayes estimator shrinks towards a point different than the sample mean in samples of small or moderate size.

The use of Monte Carlo integration methods for the hierarchical Bayes analysis was the subject of Chapter 5. We focused on Poisson/log-normal models. In addition to the vague hyperprior setting discussed in Chapter 4, we also considered a structure allowing for more prior information, with the assumption of a Inv-$\chi^2(\nu, \lambda)$ hyperprior distribution for the normal variance σ^2. The importance sampling technique proposed in Chapter 3 for the empirical Bayes analysis, was initially employed. The results demonstrated that the recommended gamma importance density is not suitable in the considered models, giving unbounded importance weights and large standard error of simulation.

We then concentrated on MCMC techniques, employing the Gibbs sampler to obtain detailed posterior inferences under a full hierarchical Bayes analysis. As the method relies on simulation from the full conditional distributions of all model parameters, we had to overcome the problem of the nonstandard form of the full conditionals of the Poisson parameters θ_i, $i = 1, \ldots, m$. Motivated by the the form of the corresponding probability density function $p(\theta_i|\mu, \sigma^2, \mathbf{y})$, we suggested an approximation based on a mixture distribution consisting of a

log-normal and a gamma component. The mixture approximation was specified to have the same first three moments as the exact full conditional distribution. This was achieved by first matching the mean and the variance of the exact and the component distributions. Then, the mixing coefficient was determined in a manner such that the skewness of the approximating distribution is equal to that of the distribution of interest. The method requires that the moments of the full conditional distribution are known. As the latter was given in nonclosed form, further approximations were needed. Although this problem was successfully addressed in Chapter 3 using the importance sampling technique, the incorporation of the method in each Gibbs sampling iteration necessitated the use of a less computationally intensive approach. We proposed a method which provides a discrete approximation to $p(\theta_i|\mu, \sigma^2, \mathbf{y})$, combining the Poisson likelihood with a discrete distribution having the same first 10 moments as the normal prior of $\log(\theta_i)$. Since the moments of interest can be obtained through the marginal density of the data, we also developed a method which relies on deriving the required marginal density by minimising the entropy distance between $p(\mathbf{y}, \boldsymbol{\gamma})$ and a suitably expressed approximation. A combination of the discretisation and the entropy-based technique provided a good approximation to the moments of the full conditional distribution, with some discrepancies occurring mainly when $y_i = 0$ and the variation of the distribution is large.

The performance of the resulting Gibbs sampling algorithm was illustrated through data examples, which demonstrated that in general the method performs well. Nevertheless, some inaccuracies are likely to occur especially when little prior information is assumed and a large number of zero counts is observed, due to the poor approximation of the full conditional moments in this case. Thus, for the exact hierarchical Bayes analysis we then employed a hybrid MCMC technique, by including a Metropolis-Hastings step within the Gibbs sampling scheme for simulating from the nonstandard full conditional distributions. Using the log-normal/gamma mixture as the proposal distribution, the Metropolis-Hastings rejection subalgorithm corrected the simulations, providing exact inferences. In most of the implementations of the method, the acceptance rate of the Metropolis-Hastings subalgorithm was outstandingly high, reaching 99% in some cases. As far as the different prior assumptions were concerned, the analysis of the examples showed that assuming largely vague hyperprior distributions at the lower stage of the hierarchical model, can result in considerably larger posterior variation and therefore wider intervals. Also, further empirical experimentation demonstrated that the shrinking direction of the hierarchical Bayes posterior mean is similar to that of the empirical Bayes estimator mentioned before, suggesting that the

direction of shrinkage cannot be determined using a fixed or adaptive function of the data. Obtaining an analytical expression for the shrinking behaviour in the considered models could be the subject of further research work.

The average risk of the hierarchical Bayes estimators was examined in Chapter 6, demonstrating superior frequency properties than the usual maximum likelihood estimator in the vast majority of the different loss functions and true parameter settings that we considered. The worst frequency properties for the hierarchical Bayes methods were observed when relatively high variation in the Poisson parameters was combined with a heavily weighted loss function. Assuming vague prior information in the analysis, provides estimators that possess very good frequency properties in a wider parameter range, in comparison to methods resulting under strongly informative modelling. The latter showed an average risk behaviour similar to that of the empirical Bayes estimators, reflecting the underlying smaller prior uncertainty. The approximate Gibbs sampling method also provided hierarchical Bayes estimators possessing low average risk, with frequency properties similar to those of the exact method.

Finally, we presented extensions to the discussed Poisson models, considering the possibility to include different exposure periods and explanatory variables in a regression structure. It was shown that through suitable modelling, the proposed methodology can be easily applied to the presented event rate and random effect generalised linear models, with the approximate Gibbs sampling approach providing good results. Future research was also suggested, which may be directed towards the implementation of similar methods in further extended situations, including more general random effects generalised linear models, possibly allowing for spatial correlation.

Appendix A
Probability distributions

In this appendix we present the main probability distributions used throughout this thesis. For each distribution we provide the notation used in the text, the density function, the range of the random variable concerned, parameter restrictions and the mean and variance of the distribution. We denote the random variable involved by X and the density functions by $f(x)$.

- **Gamma:** $X \sim \text{Ga}(a, b)$

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \qquad x > 0,\ a, b > 0,$$

$\text{E}(X) = \frac{a}{b}$, $\text{var}(X) = \frac{a}{b^2}$.

- **Chi-square:** $X \sim \chi^2_\nu$

$$f(x) = \frac{2^{-\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}, \qquad x > 0,\ \nu > 0,$$

$\text{E}(X) = \nu$, $\text{var}(X) = 2\nu$.

- **Inverse chi-square:** $X \sim \text{Inv-}\chi^2_\nu$

$$f(x) = \frac{2^{-\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} x^{-(\frac{\nu}{2}+1)} e^{-\frac{1}{2x}}, \qquad x > 0,\ \nu > 0,$$

$\text{E}(X) = \frac{1}{\nu-2}$, $(\nu > 2)$, $\text{var}(X) = \frac{2}{(\nu-2)^2(\nu-4)}$, $(\nu > 4)$.

- **Scaled inverse chi-square:** $X \sim \text{Inv-}\chi^2(\nu, \lambda)$

$$f(x) = \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} \lambda^{\frac{\nu}{2}} x^{-(\frac{\nu}{2}+1)} e^{-(\frac{\nu\lambda}{2x})}, \qquad x > 0,\ \nu, \lambda > 0,$$

$\text{E}(X) = \frac{\nu}{\nu-2}\ \lambda$, $(\nu > 2)$, $\text{var}(X) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)}\ \lambda^2$, $(\nu > 4)$.

- **Log-normal:** $X \sim \mathrm{LN}(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left\{-\frac{1}{2\sigma^2}\,(\log x - \mu)^2\right\}, \qquad -\infty < x < \infty,$$
$$-\infty < \mu < \infty,\ \sigma > 0,$$

$\mathrm{E}(X) = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$, $\mathrm{var}(X) = \mathrm{E}^2(X)\left(\mathrm{e}^{\sigma^2} - 1\right)$.

- **Multivariate normal:** $\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$f(\mathbf{X}) = (2\pi)^{-\frac{m}{2}}\,|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\,(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}, \qquad \mathbf{x} \in \mathbb{R}^m,$$

$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^{\mathrm{T}} \in \mathbb{R}^m$, $\boldsymbol{\Sigma}$ a positive definite $m \times m$ symmetric matrix,

$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu}$, $\mathrm{var}(\mathbf{X}) = \boldsymbol{\Sigma}$.

- **Poisson:** $X \sim \mathrm{Poisson}(\theta)$

$$f(x) = \frac{\theta^x e^{-\theta}}{x!}, \qquad x = 0, 1, 2, \ldots,\ \theta > 0,$$

$\mathrm{E}(X) = \mathrm{var}(X) = \theta$.

- **Negative binomial:** $X \sim \mathrm{Neg.Bin.}(\alpha, \beta)$

$$f(x) = \binom{x + \alpha - 1}{\alpha - 1}\left(\frac{\beta}{\beta + 1}\right)^{\alpha}\left(\frac{1}{\beta + 1}\right)^{x}, \qquad x = 0, 1, 2, \ldots,\ \alpha, \beta > 0,$$

$\mathrm{E}(X) = \frac{\alpha}{\beta}$, $\mathrm{var}(X) = \frac{\alpha}{\beta^2}\,(\beta + 1)$.

Appendix B
Proof of Theorem 2.1

We first make the following assumptions.

Assumption 1. For an integer $\lambda_i \geq 0$, $H_i(Y_i)$ is an arbitrary nondecreasing function such that

$$H_i(Y_i) - H_i(Y_i - 1) \geq \frac{1}{Y_i}, \qquad \text{if} \qquad Y_i > \lambda_i,$$
$$H_i(j) = 0, \qquad \text{if} \qquad j \leq 0,$$
$$\text{and} \qquad H_i(\lambda_i) = 0.$$

Assumption 2. $d_i(Y_i)$ is a nonnegative function such that $d_i(Y_i) > 0$ if $Y_i > 0$, and

$$Y_i \; H_i(Y_i - 1) \; \Delta_i d_i(Y_i) \leq \beta \; min\{d_i(Y_i - 1), d_i(Y_i)\}.$$

for all $Y_i \neq \lambda_i$ and some nonnegative constant β.

Assumption 3. There exists a finite constant K such that

$$\sum_{i=1}^{m} \frac{1}{2} \; H_i^2(Y_i) \leq K \; D(\mathbf{Y}).$$

Proof of Theorem 2.1. We notice from (2.11) that $\mathfrak{D}(\mathbf{Y})$ is expressed as the sum of two terms. Considering these two sums separately, we arrive at two inequalities which are then combined to prove the theorem. We first look at the sum $\sum_{i=1}^{m} Y_i \; \Delta_i g_i(\mathbf{Y})$. Dividing (2.14) by $D(\mathbf{Y} - \mathbf{e}_i)$, we can write

$$\frac{C(\mathbf{Y} - \mathbf{e}_i) \; H_i(Y_i - 1)}{D(\mathbf{Y} - \mathbf{e}_i)} \leq \frac{C(\mathbf{Y}) \; H_i(Y_i - 1)}{D(\mathbf{Y} - \mathbf{e}_i)}$$

and then subtracting $\frac{C(\mathbf{Y}) \; H_i(Y_i)}{D(\mathbf{Y})}$ from both sides of the above inequality we obtain

$$-\frac{C(\mathbf{Y}) \; H_i(Y_i)}{D(\mathbf{Y})} + \frac{C(\mathbf{Y} - \mathbf{e}_i) \; H_i(Y_i - 1)}{D(\mathbf{Y} - \mathbf{e}_i)} \leq -\frac{C(\mathbf{Y}) \; H_i(Y_i)}{D(\mathbf{Y})} + \frac{C(\mathbf{Y}) \; H_i(Y_i - 1)}{D(\mathbf{Y} - \mathbf{e}_i)}$$

which using the form of $g_i(\mathbf{Y})$ in (2.13) gives

$$\Leftrightarrow \Delta_i g_i(\mathbf{Y}) \leq -C(\mathbf{Y})\ \Delta_i \left\{ \frac{H_i(Y_i)}{D(\mathbf{Y})} \right\}. \tag{B.1}$$

Now, if we add and subtract the term $\frac{H_i(Y_i-1)}{D(\mathbf{Y})}$ from the difference in the right hand side of the last inequality, this can be written as

$$-\Delta_i \left\{ \frac{H_i(Y_i)}{D(\mathbf{Y})} \right\} = \frac{-H_i(Y_i) + H_i(Y_i-1)}{D(\mathbf{Y})} - \frac{H_i(Y_i-1)}{D(\mathbf{Y})} + \frac{H_i(Y_i-1)}{D(\mathbf{Y}-\mathbf{e}_i)}$$

$$= \frac{-H_i(Y_i) + H_i(Y_i-1)}{D(\mathbf{Y})} + \frac{-D(\mathbf{Y}-\mathbf{e}_i)\ H_i(Y_i-1) + D(\mathbf{Y})\ H_i(Y_i-1)}{D(\mathbf{Y})\ D(\mathbf{Y}-\mathbf{e}_i)}$$

$$= \frac{-\Delta_i H_i(Y_i)}{D(\mathbf{Y})} + \frac{H_i(Y_i-1)\ \Delta_i D(\mathbf{Y})}{D(\mathbf{Y})\ D(\mathbf{Y}-\mathbf{e}_i)}. \tag{B.2}$$

If we also notice that

$$\begin{aligned}
\Delta_i D(\mathbf{Y}) &= D(\mathbf{Y}) - D(\mathbf{Y}-\mathbf{e}_i) \\
&= \sum_{j=1}^{m} d_j(Y_j) - \sum_{\substack{j=1 \\ j\neq i}}^{m} d_j(Y_j) - d_i(Y_i-1) \\
&= d_i(Y_i) - d_i(Y_i-1) = \Delta_i d_i(Y_i),
\end{aligned}$$

it follows that, multiplying (B.1) by $Y_i \geq 0$, summing over $i = 1, \ldots, m$, and using (B.2), we obtain

$$\sum_{i=1}^{m} Y_i\ \Delta_i g_i(\mathbf{Y}) \leq \frac{C(\mathbf{Y})}{D(\mathbf{Y})} \sum_{i=1}^{m} \left\{ -Y_i\ \Delta_i H_i(Y_i) + \frac{\{Y_i H_i(Y_i-1)\ \Delta_i d_i(Y_i)\}_+}{D(\mathbf{Y}-\mathbf{e}_i)} \right\}. \tag{B.3}$$

We now notice that Assumption 1 gives

$$\begin{aligned}
Y_i\ \Delta_i H_i(Y_i) \geq 1 \Rightarrow \sum_{i=1}^{m} Y_i\ \Delta_i H_i(Y_i) &\geq m \geq N(Y) \\
\Rightarrow -\sum_{i=1}^{m} Y_i\ \Delta_i H_i(Y_i) &\leq -N(Y).
\end{aligned} \tag{B.4}$$

Therefore, setting $D'(\mathbf{Y}) = \sum_{j=1}^{m} \min\{d_j(Y_j-1), d_j(Y_j)\}$, (B.3) implies that

$$\sum_{i=1}^{m} Y_i\ \Delta_i g_i(\mathbf{Y}) \leq \frac{C(\mathbf{Y})}{D(\mathbf{Y})} \left[-N(Y) + \frac{\sum_{i=1}^{m} \{Y_i H_i(Y_i-1)\ \Delta_i d_i(Y_i)\}_+}{D'(\mathbf{Y})} \right]. \tag{B.5}$$

The transition from (B.3) to (B.5) through (B.4) implies that the inequality (B.5) will be strict for the Y_i for which $C(\mathbf{Y}) \neq 0$, and $H_i(Y_i-1)\ \Delta_i d_i(Y_i) > 0$ for at

least two i to ensure that the positive term does not coincide with the term giving $D(\mathbf{Y}-\mathbf{e}_i) = D'(\mathbf{Y})$, that is for the Y_i satisfying (2.17). Now, using Assumption 2 we obtain

$$\sum_{i=1}^{m}\{Y_i H_i(Y_i-1)\ \Delta_i d_i(Y_i)\} \leq \beta\ \sum_{i=1}^{m}\min\{d_j(Y_j-1), d_j(Y_j)\}$$

from which it follows that

$$\frac{\sum_{i=1}^{m}\{Y_i H_i(Y_i-1)\ \Delta_i d_i(Y_i)\}}{D'(\mathbf{Y})} \leq \beta.$$

Then, employing the last inequality, (B.5) becomes

$$\sum_{i=1}^{m} Y_i\ \Delta_i g_i(\mathbf{Y}) \leq \frac{C(\mathbf{Y})}{D(\mathbf{Y})}\ \{\beta - N(Y)\}. \tag{B.6}$$

We now consider the second sum in (2.11). From Assumption 3, if we multiply both sides of the inequality by $\frac{C^2(\mathbf{Y})}{D^2(\mathbf{Y})}$ we have that

$$\sum_{i=1}^{m}\frac{1}{2}\ \frac{C^2(\mathbf{Y})\ H_i^2(Y_i)}{D^2(\mathbf{Y})} \leq \frac{K\ C^2(\mathbf{Y})}{D(\mathbf{Y})}$$

and using (2.13), the above implies that

$$\sum_{i=1}^{m}\frac{1}{2}\ g_i^2(\mathbf{Y}) \leq \frac{K\ C^2(\mathbf{Y})}{D(\mathbf{Y})}\,. \tag{B.7}$$

Adding together (B.6) and (B.7) and taking into account (2.11) we obtain

$$\mathfrak{D}(\mathbf{Y}) \leq -\frac{C(\mathbf{Y})}{D(\mathbf{Y})}\ \{N(Y) - \beta - K\ C(\mathbf{Y})\}. \tag{B.8}$$

However, from condition (2.15) we have that

$$-C(\mathbf{Y})\ \{N(Y) - \beta - K\ C(\mathbf{Y})\} \leq 0$$

and therefore, (B.8) gives

$$\mathfrak{D}(\mathbf{Y}) \leq 0.$$

Appendix C
Vector and matrix differentiation

In this appendix we provide some calculus results regarding vector and matrix differentiation (e.g. see Magnus and Neudecker, 1988). We let $\mathbf{x}$ and $\mathbf{a}$ be $m \times 1$ vectors, and $\mathbf{X}$ and $\mathbf{A}$ be $m \times m$ matrices. We also let $f(\cdot)$ denote a real function of $\mathbf{x}$ or $\mathbf{X}$.

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1} \quad \frac{\partial f(\mathbf{x})}{\partial x_2} \quad \cdots \quad \frac{\partial f(\mathbf{x})}{\partial x_m} \right)^{\mathrm{T}}$$

$$\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \begin{pmatrix} \frac{\partial f(\mathbf{X})}{\partial x_{11}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{1m}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{X})}{\partial x_{m1}} & \cdots & \frac{\partial f(\mathbf{X})}{\partial x_{mm}} \end{pmatrix}$$

$$\frac{\partial (\mathbf{a}^{\mathrm{T}}\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial (\mathbf{x}^{\mathrm{T}}\mathbf{a})}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial (\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x})}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

$$\frac{\partial \{\mathrm{tr}(\mathbf{X})\}}{\partial \mathbf{X}} = \mathbf{I}_m$$

$$\frac{\partial \{\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{X})\}}{\partial \mathbf{X}} = 2\mathbf{X}$$

$$\frac{\partial \{\mathrm{tr}(\mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{X})\}}{\partial \mathbf{X}} = 2\mathbf{A}\mathbf{X}$$

$$\frac{\partial \{\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{A}\mathbf{X})\}}{\partial \mathbf{X}} = \left(\mathbf{A} + \mathbf{A}^{\mathrm{T}}\right)\mathbf{X}$$

$$\frac{\partial (\log|\mathbf{X}|)}{\partial \mathbf{X}} = \left(\mathbf{X}^{\mathrm{T}}\right)^{-1}$$

References

Aitchison, J. and Brown, J. (1957). *The Lognormal Distribution.* Cambridge: Cambridge University Press.

Albert, J.H. (1981). Simultaneous estimation of Poisson means. *J. of Multivariate Analysis*, **11**, 400–417.

Albert, J.H. (1985). Simultaneous estimation of Poisson means under exchangeable and independence models. *J. Statist. Comput. Simul.*, **23**, 1–14.

Albert, J.H. (1988). Bayesian estimation of Poisson means using a hierarchical log-linear model. In *Bayesian Statistics 3*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., New York: Oxford University Press, pp. 519–531.

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

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B, **36**, 192–236.

Best, N.G., Cowles, M.K. and Vines, K. (1995). CODA: Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30. Technical report, MRC Biostatistics Unit, Cambridge.

Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis.* New York: Wiley.

Breslow, N.E. and Clayton, D.G. (1993). Inference in generalized linear mixed models. *J. Amer. Statist. Assoc.*, **88**, 9–25.

Carlin, B.P. and Gelfand, A.E. (1991). An iterative Monte Carlo method for nonconjugate Bayesian analysis. *Statistics and Computing*, **1**, 119–128.

Carlin, B.P. and Louis, T.A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis.* London: Chapman and Hall.

Casella, G. and George, E. (1992). Explaining the Gibbs sampler. *Amer. Statistician*, **46**, 167–174.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *Amer. Statistician*, **59**, 327–335.

Christiansen, C.L. and Morris, C.N. (1996). Fitting and checking a two-level Poisson model: Modeling patient mortality rates in heart transplant patients. In *Bayesian Biostatistics*, D. Berry and D. Stangl, eds., New York: Marcel Dekker, pp. 467–501.

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

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

Clevenson, M.L. and Zidek, J.V. (1975). Simultaneous estimation of the means of independent Poisson laws. *J. Amer. Statist. Assoc.*, **70**, 698–705.

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

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

Daniels, M.J. (1999). A prior for the variance in hierarchical models. *Canadian J. Statist.*, **27**, 567–578.

Deely, J.J. and Lindley, D.V. (1981). Bayes empirical Bayes. *J. Amer. Statist. Assoc.*, **76**, 833–841.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New York: Springer-Verlag.

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

Ericson, W.A. (1969). A note on the posterior mean of a population mean. *J. R. Statist. Soc.* B, **31**, 332–334.

Gaver, D.P. and O'Muircheartaigh, I.G. (1987). Robust empirical Bayes analysis of event rates. *Technometrics*, **29**, 1–15.

Gelfand, A.E., Hills, S.I., Racine-Poon, A. and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models. *J. Amer. Statist. Assoc.*, **90**, 972–985.

Gelfand, A.E. and 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. and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

George, E.I., Makov, U.E. and Smith, A.F.M. (1993). Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, **20**, 147–156.

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

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

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4*, J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 169–193.

Geyer, C.J. (1992). Practical Markov chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473–511.

Ghosh, M., Hwang, J.T. and Tsui, K.W. (1983). Construction of improved estimators in multiparameter estimation for discrete exponential families. *Ann. Statist.*, **11**, 351–367.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., eds. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Gilks, W.R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Appl. Statist.*, **41**, 337–348.

Goldstein, M. (1975). A note on some Bayesian nonparametric estimates. *Ann. Statist.*, **3**, 736–740.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hodges, J.L., JR, and Lehmann, E.L. (1951). Some applications of the Cramer-Rao inequality. In *Proc. Second Berkeley Symp. on Math. Statist. and Prob.*, pp. 13–20.

Hudson, H.M. (1974). Empirical Bayes estimation. Technical report No. 58, Department of Statistics, Stanford University.

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

Hudson, H.M. (1985). Adaptive estimators for simultaneous estimation of Poisson means. *Ann. Statist.*, **13**, 246–261.

Hudson, H.M. and Tsui, K.W. (1981). Simultaneous Poisson estimators for a priori hypotheses about means. *J. Amer. Statist. Assoc.*, **76**, 182–187.

Hwang, J.T. (1982). Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases. *Ann. Statist.*, **10**, 857–867.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, **1**, Berkeley: University of California Press, pp. 361–379.

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions, Volume 1.* (2nd edition). New York: Wiley.

Kass, R.E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *J. Amer. Statist. Assoc.*, **84**, 717–726.

Leonard, T. (1976). Some alternative approaches to multiparameter estimation. *Biometrika*, **63**, 69–75.

Leonard, T. (1977). A Bayesian approach to some multinomial estimation and pretesting problems. *J. Amer. Statist. Assoc.*, **72**, 869–874.

Leonard, T. and Novick, M.R. (1986). Bayesian full rank marginalization for two-way contingency tables. *J. Educ. Behav. Statist.*, **11**, 33–56.

Lindley, D.V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Stat.*, **27**, 986–1005.

Magnus, J.R. and Neudecker, H. (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics.* Chichester: Wiley.

Matsumura, E. and Tsui, K. (1982). Stein-type Poisson estimators in audit sampling. *J. Accounting Research*, **20**, 162–170.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics*, **21**, 1087–1091.

Morris, C.N. (1983a). Parametric empirical Bayes inference: Theory and applications. *J. Amer. Statist. Assoc.*, **78**, 47–65.

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

Muller, P. (1991). Metropolis based posterior integration schemes. Technical report, Statistics Department, Purdue University.

Peng, J.C.M. (1975). Simultaneous estimation of the parameters of independent Poisson distributions. Technical report No. 78, Department of Statistics, Stanford University.

Proschan, F. (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, **5**, 375–383.

Raftery, A.E. and Lewis, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics 4*, J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith, eds., Oxford: Oxford University Press, pp. 763–773.

Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy Gibbs sampler. *J. Amer. Statist. Assoc.*, **87**, 862–868.

Roberts, G.O. (1995). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds., London: Chapman and Hall, pp. 45–57.

Rosenkrantz, R.D. (1989). *E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, Dordrecht: Kluwer Academic.

Smith, A.F.M. and Roberts, G.O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 1–24.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1995). BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.5. Technical report, MRC Biostatistics Unit, Cambridge.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. on Math. Statist. and Prob.*, **1**, Berkeley: University of California Press, pp. 197–206.

Stein, C. (1973). Estimation of the mean of a multivariate distribution. *Proc. Prague Symp. Asymp. Statist.* 345–381.

Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory.* (6th edition). London: Edward Arnold.

Tempelman, R.J. and Gianola, D. (1993). Marginal maximum likelihood estimation of variance components in Poisson mixed models using Laplacian integration. *Genetics, Selection, Evolution*, **25**, 305–319.

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

Tierney, L. (1995). Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*, W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds., London: Chapman and Hall, pp. 59–74.

Tsui, K.W. (1981). Simultaneous estimation of several Poisson parameters under squared error loss. *Ann. Inst. Statist. Math.*, **33**, 215–223.

Tsui, K.W. and Press, S.J. (1982). Simultaneous estimation of several Poisson parameters under K-normalized squared error loss. *Ann. Statist.*, **10**, 93–100.

Wakefield, J.C., Gelfand, A.E. and Smith, A.F.M. (1991). Efficient generation of random variates via the ratio-of-uniform method. *Statistics and Computing*, **1**, 129–133.

Worledge, D.H., Strigham, R.S. and McClymont, A.S. (1982). PWR power plant reliability data. Interim report NP-2592, Electric Power Research Institute.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Amer. Statist. Assoc.*, **86**, 79–86.

Zellner, A. and Min, C.K. (1995). Gibbs sampler convergence criteria. *J. Amer. Statist. Assoc.*, **90**, 921–927.