# AN INTRODUCTION TO MARKOV CHAIN MONTE CARLO METHODS AND THEIR ACTUARIAL APPLICATIONS

DAVID P. M. SCOLLNIK

Department of Mathematics and Statistics
University of Calgary

*Abstract*

*This paper introduces the readers of the* Proceedings *to an important class of computer based simulation techniques known as Markov chain Monte Carlo (MCMC) methods. General properties characterizing these methods will be discussed, but the main emphasis will be placed on one MCMC method known as the Gibbs sampler. The Gibbs sampler permits one to simulate realizations from complicated stochastic models in high dimensions by making use of the model's associated full conditional distributions, which will generally have a much simpler and more manageable form. In its most extreme version, the Gibbs sampler reduces the analysis of a complicated multivariate stochastic model to the consideration of that model's associated univariate full conditional distributions.*

*In this paper, the Gibbs sampler will be illustrated with four examples. The first three of these examples serve as rather elementary yet instructive applications of the Gibbs sampler. The fourth example describes a reasonably sophisticated application of the Gibbs sampler in the important arena of credibility for classification ratemaking via hierarchical models, and involves the Bayesian prediction of frequency counts in workers compensation insurance.*

## 1. INTRODUCTION

The purpose of this paper is to acquaint the readership of the *Proceedings* with a class of simulation techniques known as Markov chain Monte Carlo (MCMC) methods. These methods permit a practitioner to simulate a dependent sequence of random draws from very complicated stochastic models. The main emphasis will be placed on one MCMC method known as the Gibbs sampler. It is not an understatement to say that several hundred papers relating to the Gibbs sampling methodology have appeared in the statistical literature since 1990. Yet, the Gibbs sampler has made only a handful of appearances within the actuarial literature to date. Carlin [3] used the Gibbs sampler in order to study the Bayesian state-space modeling of non-standard actuarial time series, and Carlin [4] used it to develop various Bayesian approaches to graduation. Klugman and Carlin [19] also used the Gibbs sampler in the arena of Bayesian graduation, this time concentrating on a hierarchical version of Whittaker-Henderson graduation. Scollnik [24] studied a simultaneous equations model for insurance ratemaking, and conducted a Bayesian analysis of this model with the Gibbs sampler.

This paper reviews the essential nature of the Gibbs sampling algorithm and illustrates its application with four examples of varying complexity. This paper is primarily expository, although references are provided to important theoretical results in the published literature. The reader is presumed to possess at least a passing familiarity with the material relating to statistical computing and stochastic simulation present in the syllabus for CAS Associateship Examination Part 4B. The theoretical content of the paper is mainly concentrated in Section 2, which provides a brief discussion of Markov chains and the properties of MCMC methods. Except for noting Equations 2.1 and 2.2 along with their interpretation, the reader may skip over Section 2 the first time through reading this paper. Section 3 formally introduces the Gibbs sampler and illustrates it with an example. Section 4 discusses some of the practical considerations related to the implementation of a Gibbs sampler. In Section 5, some aspects of Bayesian inference using Gibbs sampling are considered, and two final examples are presented. The first of these concerns the Bayesian estimation of the parameter for a size of loss distribution when grouped data are observed. The second addresses credibility for classification ratemaking via hierarchical models and involves the Bayesian prediction of frequency counts in workers compensation insurance. In Section 6 we conclude our presentation and point out some areas of application to be explored in the future.

Since the subject of MCMC methods is still foreign to most actuaries at this time, we will conclude this section with a simple introductory example, which we will return to in Section 3.

*Example 1*

This example starts by recalling that a generalized Pareto distribution can be constructed by mixing one gamma distribution with another gamma distribution in a certain manner. (See for example, Hogg and Klugman [16, pp. 53–54].) More precisely, if a loss random variable $X$ has a conditional gamma $(k,\theta)$ distribution with density

$$f(x \mid \theta) = \frac{\theta^k}{\Gamma(k)} x^{k-1} \exp(-\theta x), \qquad 0 < x < \infty, \qquad (1.1)$$

and the mixing random variable $\theta$ has a marginal gamma $(\alpha,\lambda)$ distribution with density

$$f(\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp(-\lambda\theta), \qquad 0 < \theta < \infty,$$

then $X$ has a marginal generalized Pareto $(\alpha,\lambda,k)$ distribution with density

$$f(x) = \frac{\Gamma(\alpha+k)\lambda^\alpha x^{k-1}}{\Gamma(\alpha)\Gamma(k)(\lambda+x)^{\alpha+k}}, \qquad 0 < x < \infty.$$

It also follows that the conditional distribution of $\theta$ given $X$ is also given by a gamma distribution, namely,

$$f(\theta \mid x) \sim \text{gamma } (\alpha + k, \lambda + x), \qquad 0 < \theta < \infty. \qquad (1.2)$$

We will perform the following iterative sampling algorithm, which is based upon the conditional distributions appearing in Equations 1.1 and 1.2:

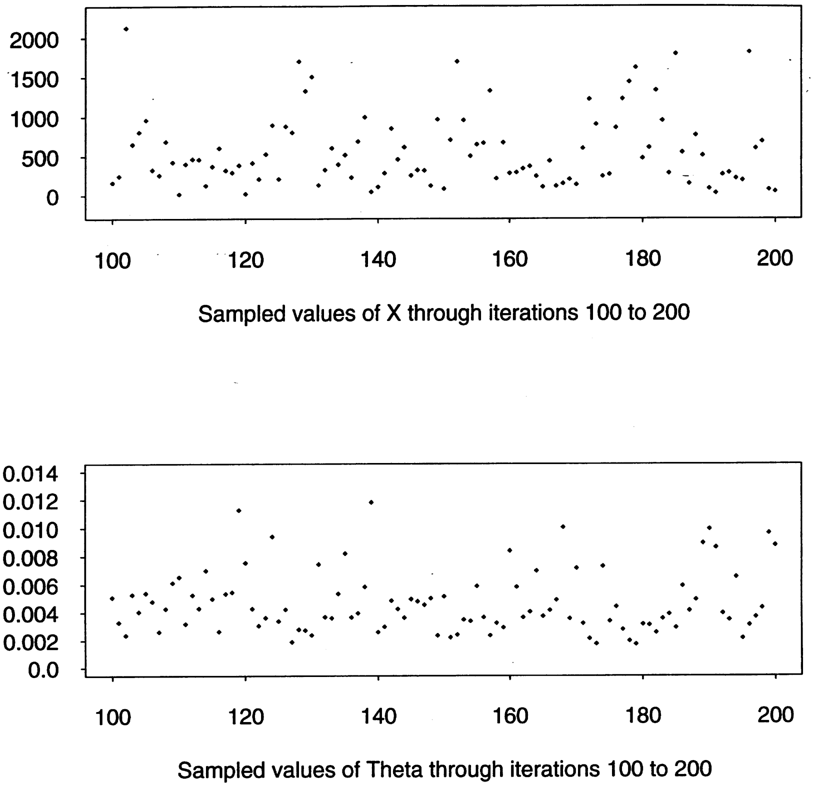
1. Select arbitrary starting values $X^{(0)}$ and $\theta^{(0)}$.
2. Set the counter index $i = 0$.
3. Sample $X^{(i+1)}$ from $f(x \mid \theta^{(i)}) \sim \text{gamma } (k, \theta^{(i)})$.
4. Sample $\theta^{(i+1)}$ from $f(\theta \mid X^{(i+1)}) \sim \text{gamma } (\alpha + k, \lambda + X^{(i+1)})$.
5. Set $i \leftarrow i + 1$ and return to Step 3.

For the sake of illustration, we assigned the model parameters $\alpha = 5$, $\lambda = 1000$ and $k = 2$ so that the marginal distribution of $\theta$ is gamma $(5, 1000)$ with mean 0.005 and the marginal distribution of $X$ is generalized Pareto $(5, 1000, 2)$ with mean 500. We then ran the algorithm described above on a fast computer for a total of 500 iterations and stored the sequence of generated values $X^{(0)}, \theta^{(0)}, X^{(1)}, \theta^{(1)}, \ldots, X^{(499)}, \theta^{(499)}, X^{(500)}, \theta^{(500)}$. It must be emphasized that this sequence of random draws is clearly not independent, since $X^{(1)}$ depends upon $\theta^{(0)}$, $\theta^{(1)}$ depends upon $X^{(1)}$, and so forth. Our two starting values were arbitrarily selected to be $X^{(0)} = 20$ and $\theta^{(0)} = 10$. The sequence of sampled values for $X^{(i)}$ is plotted in Figure 1, along with the sequence of sampled values for $\theta^{(i)}$, for iterations 100 through 200. Both sequences do appear to be random, and some dependencies between successive values are discernible in places.

In Figure 2, we plot the histograms of the last 500 values appearing in each of the two sequences of sampled values (the starting values $X^{(0)}$ and $\theta^{(0)}$ were discarded at this point).

FIGURE 1

SAMPLE PATHS FOR $X^{(i)}$ AND $\theta^{(i)}$ IN EXAMPLE 1

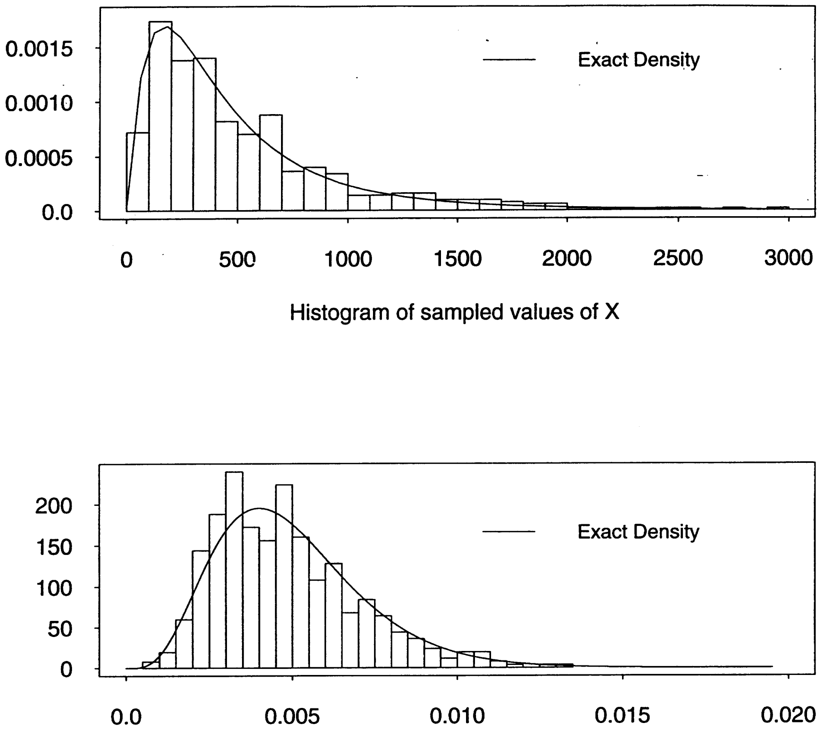


In each plot, we also overlay the actual density curve for the marginal distribution of either $X$ or $\theta$. Surprisingly, the dependent sampling scheme we implemented, which was based upon the full conditional distributions $f(\theta \mid x)$ and $f(x \mid \theta)$, appears to have generated random samples from the underlying marginal distributions.

Now, notice that the marginal distribution of $X$ may be interpreted as the average of the conditional distribution of $X$ given

FIGURE 2

HISTOGRAMS OF SAMPLE VALUES FOR $X$ AND $\theta$ IN EXAMPLE 1



$\theta$ taken with respect to the marginal distribution of $\theta$; that is,

$$f(x) = \int f(x \mid \theta) f(\theta) d\theta.$$

Since the sampled values of $\theta^{(i)}$ appear to constitute a random sample of sorts from the marginal distribution of $\theta$, this suggests that a naive estimate of the value of the marginal density function for $X$ at the point $x$ might be constructed by taking the empirical average of $f(x \mid \theta^{(i)})$ over the sampled values for $\theta^{(i)}$. If $\theta^{(1)} =$

0.0055, for example, then

$$f(x \mid \theta^{(1)}) = 0.0055^2 x \exp(-0.0055x).$$

One does a similar computation for the other values of $\theta^{(i)}$ and averages to get

$$\hat{f}(x) = \frac{1}{500} \sum_{i=1}^{500} f(x \mid \theta^{(i)}). \tag{1.3}$$

Similarly, we might construct

$$\hat{f}(\theta) = \frac{1}{500} \sum_{i=1}^{500} f(\theta \mid X^{(i)}) \tag{1.4}$$

as a density estimate of $f(\theta)$. These estimated density functions are plotted in Figure 3 along with their exact counterparts, and it is evident that the estimated densities happen to be excellent.

## 2. MARKOV CHAIN MONTE CARLO

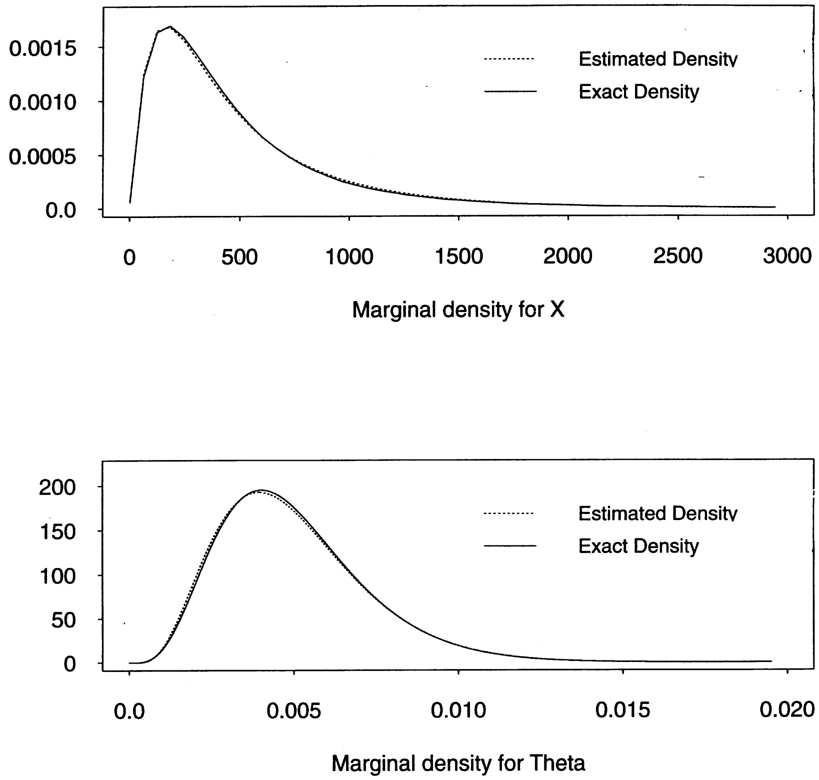
In the example of the previous section, we considered an iterative simulation scheme that generated two dependent sequences of random variates. Apparently, we were able to use these sequences in order to capture characteristics of the underlying joint distribution that defined the simulation scheme in the first place. In this section, we will discuss a few properties of certain simulation schemes that generate dependent sequences of random variates and note in what manner these dependent sequences may be used for making useful statistical inference. The main results are given by Equations 2.1 and 2.2.

Before we begin, it may prove useful to quickly, and very informally, review some elementary Markov chain theory. Tierney [31] provides a much more detailed and rigorous discussion of this material. A Markov chain is just a collection of random variables $\{X_n;\ n \geq 0\}$, with the distribution of the random variables on some space › $\subseteq R^k$ governed by the transition probabilities

$$\Pr(X_{n+1} \in A \mid X_0, \ldots, X_n) = K(X_n, A),$$

FIGURE 3

ESTIMATED AND EXACT MARGINAL DENSITIES FOR $X$ AND $\theta$ IN EXAMPLE 1



where $A \subset$ › . Notice that the probability distribution of the next random variable in the sequence, given the current and past states, depends only upon the current state. This is known as the Markov property. The distribution of $X_0$ is known as the initial distribution of the Markov chain. The conditional distribution of $X_n$ given $X_0$ is described by

$$\Pr(X_n \in A \mid X_0) = K^n(X_0, A),$$

where $K^n$ denotes the $n$th application of $K$. An invariant distribution $\pi(x)$ for the Markov chain is a density satisfying

$$\pi(A) = \int K(x, A)\, \pi(x)\, dx,$$

and it is also an equilibrium distribution if

$$\lim_{n\to\infty} K^n(x, A) = \pi(A).$$

For simplicity, we are using the notation $\pi(x)$ to identify both the distribution and density for a random variable, trusting the precise meaning to be evident from the context. A Markov chain with invariant distribution $\pi(x)$ is irreducible if it has a positive probability of entering any state assigned positive probability by $\pi(x)$, regardless of the initial state or value of $X_0$. A chain is periodic if it can take on certain values only at regularly spaced intervals, and is aperiodic otherwise. If a Markov chain with a proper invariant distribution is both irreducible and aperiodic, then the invariant distribution is unique and it is also the equilibrium distribution of the chain.

A MCMC method is a sampling based simulation technique that may be used in order to generate a dependent sample from a certain distribution of interest. Formally, a MCMC method proceeds by first specifying an irreducible and aperiodic Markov chain with a unique invariant distribution $\pi(x)$ equal to the desired distribution of interest (or target distribution). Curiously, there are usually a number of easy ways in which to construct such a Markov chain. The next step is to simulate one or more realizations of this Markov chain on a fast computer. Each path of simulated values will form a dependent random sample from the distribution of interest, provided that certain regularity conditions are satisfied. Then these dependent sample paths may be utilized for inferential purposes in a variety of ways. In particular, if the Markov chain is aperiodic and irreducible, with unique invariant distribution $\pi(x)$, and $X^{(1)}, X^{(2)}, \ldots,$ is a realization of this chain, then known asymptotic results (e.g., Tierney [31] or Roberts and Smith [23]) tell us that:

$$X^{(t)} \xrightarrow{d} X \sim \pi(x) \qquad \text{as} \quad t \to \infty, \tag{2.1}$$

and

$$\frac{1}{t}\sum_{i=1}^{t} h(X^{(i)}) \to \mathrm{E}_{\pi}[h(X)] \qquad \text{as} \quad t \to \infty, \textit{ almost surely.} \tag{2.2}$$

Equation 2.1 indicates that as $t$ becomes moderately large, the value $X^{(t)}$ is very nearly a random draw from the distribution of interest. In practice, a value of $t \approx 10$ to 15 is often more than sufficient. This result also allows us to generate an approximately independent random sample from the distribution with density $f(x)$ by using only every $k$th value appearing in the sequence. The value of $k$ should be taken to be large enough so that the sample autocorrelation function coefficients for the values appearing in the subsequence are reminiscent of those for a purely random process or a stochastically independent sequence, that is, until there are no significant autocorrelations at non-zero lags. This idea is illustrated in Example 2. Autocorrelation functions are covered in some depth in the course of reading for Associateship Examination Part 3A, *Applied Statistical Methods* (also see Miller and Wichern [21, pp. 333–337, 356–365]).

Equation 2.2 tells us that if $h$ is an arbitrary $\pi$-integrable real-valued function of $X$, then the average of this function taken over the realized values of $X^{(t)}$ (the ergodic average of the function) converges (almost surely, as $t \to \infty$) to its expected value under the target density. In practice, usually the first 10 to 100 values of the simulation are discarded, in order to reduce the dependence of these estimates upon the selected starting values.

Notice that if $h(X)$ is taken to be the conditional density for some random variable $Y$ given $X$, then Equation 2.2 suggests that the marginal density of $Y$ may be estimated at the point $y$ by averaging the conditional density $f(y \mid X)$ over the realized values $X^{(t)}$ (as in Gelfand and Smith [9, pp. 402–403]).

At this point, the reader is probably wondering how one would go about constructing a suitable Markov chain when a certain target density $\pi(x)$ is of interest. The so-called Gibbs sampler, a special kind of MCMC method, is one easy and very popular approach. The Gibbs sampler was introduced by Geman and Geman [11] in the context of image restoration, and its suitability for a wide range of problems in the field of Bayesian inference was recognized by Gelfand and Smith [9]. An elementary introduction to the Gibbs sampler is given in Casella and George [5], and those readers unfamiliar with the methodology are certainly encouraged to consult this reference. More sophisticated discussions of the Gibbs sampler and MCMC methods in general are given in Smith and Roberts [25], Tanner [29], and Tierney [31].

## 3. THE GIBBS SAMPLER

In order to formally introduce the Gibbs sampler, let us begin by letting the target distribution $\pi(x)$ now correspond to a joint distribution $\pi(x_1, x_2, \ldots, x_k)$. We assume that this joint distribution exists and is proper. Each of the $x_i$ terms may represent either a single random variable or, more generally, a block of several random variables grouped together. Let $\pi(x_j)$ represent the marginal distribution of the $j$th block of variables, $x_j$, and let $\pi(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$ represent the full conditional distribution of the $j$th block of variables, given the remainder. Besag [2] observed that the collection of full conditional distributions uniquely determines the joint distribution, provided that the joint distribution is proper. The Gibbs sampler utilizes a set of full conditional distributions associated with the target distribution of interest in order to define a Markov chain with an invariant distribution equal to the target distribution. When we speak of a Gibbs sampler, we are actually referring to an implementation of the following iterative sampling scheme:

1. Select initial values $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, \ldots, x_k^{(0)})$.
2. Set the counter index $i = 0$.
3. Simulate the sequence of random draws:

$$\begin{aligned} x_1^{(i+1)} &\sim \pi(x_1 \mid x_2^{(i)}, \ldots, x_k^{(i)}), \\ x_2^{(i+1)} &\sim \pi(x_2 \mid x_1^{(i+1)}, x_3^{(i)}, \ldots, x_k^{(i)}), \\ x_3^{(i+1)} &\sim \pi(x_3 \mid x_1^{(i+1)}, x_2^{(i+1)}, x_4^{(i)}, \ldots, x_k^{(i)}), \\ &\vdots \\ x_k^{(i+1)} &\sim \pi(x_k \mid x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_{k-1}^{(i+1)}), \end{aligned}$$

   and form

$$x^{(i+1)} = (x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_k^{(i+1)}).$$

4. Set $i \leftarrow i + 1$ and return to Step 3.

Notice that in Step 3 of the Gibbs sampling algorithm, we are required to sample random draws once from each of the full conditional distributions and that the values of the conditioning variables are sequentially updated, one by one. This sampling algorithm defines a valid MCMC method, and by its construction also ensures that the target distribution $\pi(x)$ is an invariant distribution of the Markov chain so defined (e.g., Tierney [31]). Mild regularity conditions (typically satisfied in practice) guarantee that Equations 2.1 and 2.2 will apply. Refer to Theorem 2 in Roberts and Smith [23] for one set of sufficient conditions. Notice that since Equation 2.1 implies that $x^{(i)}$ is very nearly a random draw from the joint distribution $\pi(x)$, it is also the case that each component $x_j^{(i)}$ is very nearly a random draw from the marginal distribution $\pi(x_j)$, for $j = 1,2,\ldots,k$ (provided throughout that $i$ is sufficiently large). This is a useful result to note when simulation based inference is sought with respect to one or more of the marginal distributions.

Besag [2] observed the fact that the collection of full conditional distributions uniquely determines the joint distribution, provided that the joint distribution exists and is proper. However, it is not the case that a collection of proper full conditional distributions necessarily guarantees the *existence* of a proper joint distribution for the random variables involved. For example, note that

$$f(x_1,x_2) \propto \exp(-[x_1 + x_2]^2/2),$$

with $-\infty < x_1 < \infty$ and $-\infty < x_2 < \infty$, defines an improper joint distribution with two proper univariate normal full conditional distributions (Gelfand [8]). When a set of proper full conditional distributions fails to determine a proper joint distribution, any application of the Gibbs sampling algorithm to these full conditional distributions is to be avoided. If the Gibbs sampler was invoked under these circumstances, the algorithm may either fail to converge or else converge to a state that is not readily interpretable.

By now, perhaps the reader has noticed that the example presented in Section 1 really just amounted to an application of the Gibbs sampling algorithm to the two full conditional distributions $f(x \mid \theta)$ and $f(\theta \mid x)$ appearing in Equations 1.1 and 1.2. By construction, we ensured that the joint distribution $f(x,\theta)$ also existed as a proper distribution. From the discussion above, it follows that Equation 2.1 explains why the sequence of sampled values for $X^{(i)}$ and $\theta^{(i)}$ effectively constituted random samples from the marginal distributions of $X$ and $\theta$, respectively. Similarly, the two density estimates defined by Equations 1.3 and 1.4 performed as well as they did because of the result described by Equation 2.2.

We conclude this section with a second example, before discussing some of the practical issues relating to the implementation of a Gibbs sampler in Section 4.

*Example 2*

Consider the following distributional model:

$$\begin{aligned} f(y) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \qquad 0 \le y \le 1 \\ &\sim \text{beta } (\alpha,\beta); \qquad (3.1) \\ f(n) &= [\exp(\lambda) - 1]^{-1}\frac{\lambda^n}{n!}, \qquad n = 1,2,\ldots \\ &\sim \text{zero-truncated Poisson } (\lambda); \qquad (3.2) \\ f(x \mid y,n) &= \binom{n}{x} y^x(1-y)^{n-x}, \qquad x = 0,1,\ldots,n \\ &\sim \text{binomial } (n,y). \qquad (3.3) \end{aligned}$$

We will assume that the random variables $Y$ and $N$ are independent, so that the proper joint distribution of $X$, $Y$, and $Z$ obviously exists as the product of Equations 3.1, 3.2, and 3.3. In order to give the model above an actuarial interpretation, imagine that, conditional upon $Y$ and $N$, the random variable $X$ represents the number of policies generating a claim in a portfolio of $N$ identical and independent policies, each with a claim probability equal to $Y$. A portfolio is characterized by the value of the parameters $Y$ and $N$, which are random variables in their own right with independent beta and zero-truncated Poisson distributions, respectively. The marginal distribution of $X$ describes the typical number of policies generating a claim in an arbitrary portfolio. Unfortunately, the marginal distribution of $X$ cannot be obtained in a closed form. (The reader is invited to try.) In order to study the marginal distribution of $X$, we will consider an application of the Gibbs sampler.

For the model above, the following set of full conditional distributions may be derived in a straightforward fashion:

$$f(x \mid y,n) \sim \text{binomial } (n,y); \tag{3.4}$$

$$f(y \mid x,n) \sim \text{beta } (x+\alpha, n-x+\beta); \tag{3.5}$$

$$f(n \mid x,y) = \exp(-\lambda[1-y])\frac{(\lambda[1-y])^{n-x}}{(n-x)!}, \qquad n = x, x+1, x+2, \ldots \tag{3.6}$$

or

$$f(n-x \mid x,y) \sim \text{Poisson } (\lambda[1-y]).$$

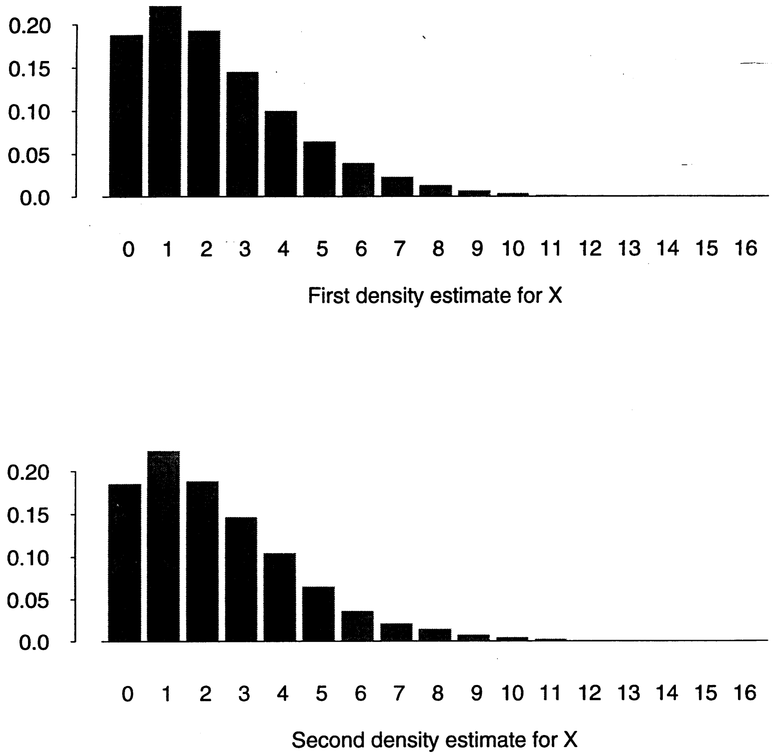
For the purpose of illustration, we set the model parameters equal to $\alpha = 2$, $\beta = 8$, and $\lambda = 12$, and initiated 5100 iterations of the Gibbs sampler using the full conditional distributions found in Equations 3.4, 3.5, and 3.6, with initial values $X^{(0)} = 4$, $Y^{(0)} = 0.5$, and $N^{(0)} = 50$. By averaging Equation 3.4 over the simulated values of $Y^{(i)}$ and $N^{(i)}$ in the spirit of Equation 2.2, after first discarding the initial 100 values in each sample path in order to 'burn-in' the Gibbs sampler and remove the effect of the starting values, a density estimate for the random variable $X$ at the point $x$ is given by the average of 5000 binomial distributions:

$$\hat{f}(x) = \frac{1}{5000}\sum_{i=101}^{5100} f(x \mid Y^{(i)}, N^{(i)}). \tag{3.7}$$

A plot of this density estimate appears in the upper half of Figure 4. For comparison, we also constructed a histogram estimate of the density for the random variable $X$ on the basis of 1000 approximately independent realizations of this random variable. These 1000 approximately independent random draws were obtained by taking or accepting every fifth of the last 5000 values for $X$ appearing in the simulation. (See the discussion in the next paragraph.) The resulting histogram density estimate appears as

FIGURE 4

TWO ESTIMATED DENSITIES FOR $X$ IN EXAMPLE 2



First density estimate for X

Second density estimate for X

the second plot in Figure 4, and we observe that it is consistent with the first estimate.

As previously mentioned in Section 2, thinning the sequence of simulated values output by a Gibbs sampler by accepting only every $k$th generated value reduces the serial correlation between the accepted values, and sample autocorrelation functions may be examined in order to assess the dependence in the thinned sequence (Miller and Wichern [21]). We applied this idea in the paragraph above to the last 5000 of the simulated values for $X$

output by the Gibbs sampler using $k = 5$, so that only every fifth of these generated values was accepted and so that 1000 values were accepted in total. The sample autocorrelation function for the original sequence of 5000 simulated values appears as the first plot in Figure 5. The heights of the twenty different spikes in this plot represent the values of the sample autocorrelation coefficients at lags 0 through 19 for this sequence of 5000 values, respectively. If this sequence of 5000 simulated values were independent, then all of the sample autocorrelations at non-zero lags should be close to zero. Spikes crossing either of the two horizontal dashed lines identify autocorrelation coefficients that are significantly different from zero (at the 95 percent level of significance). For this sequence of 5000 simulated values, we may observe that significant autocorrelations are identified at the non-zero lags 1 through 5, clearly demonstrating the dependent nature of this sequence. The sample autocorrelation function for the thinned sequence of 1000 simulated values appears as the second plot in Figure 5. This sample autocorrelation function is reminiscent of the function we would expect for a purely random process, since none of the autocorrelations at non-zero lags is significantly different from zero. This demonstrates that by thinning the original sequence of simulated values for $X$, we have indeed recovered an approximately independent random sample as claimed.

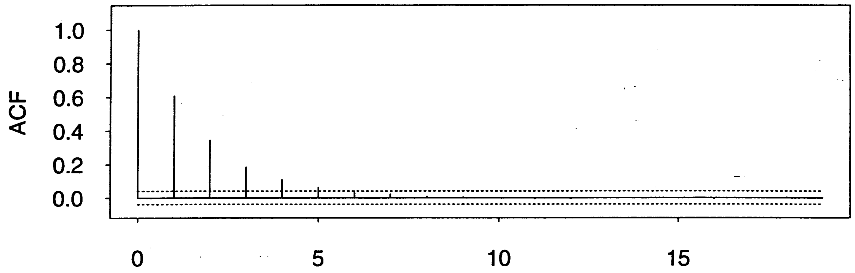
## 4. PRACTICAL CONSIDERATIONS RELATED TO GIBBS SAMPLING

There is one very simple and overriding reason for the recent popularity of the Gibbs sampler as a tool for statistical inference: it permits the analysis of any statistical model possessing a complicated multivariate distribution to be reduced to the analysis of its much simpler, and lower dimensional, full conditional distributions. In fact, all that is required is that we be able to iteratively sample a large number of random variates from these conditional distributions. Since

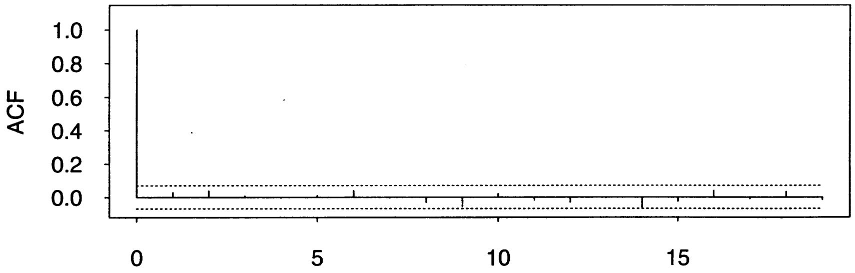
$$\pi(x_j \mid x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_k) \propto \pi(x), \tag{4.1}$$

FIGURE 5

TWO SAMPLE AUTOCORRELATION FUNCTIONS FOR $X$ IN EXAMPLE 2



Sample ACF for the last 5000 values of X in the original sequence



Sample ACF for the last 1000 values of X in the subsequence

where $\pi(x)$ on the right-hand side is viewed as a function of $x_j$ with all of the other arguments held fixed, we will always have the form of the full conditional distributions required to implement a Gibbs sampler immediately available (at least up to their normalizing constants) whenever the form of the target distribution is known. When a full conditional distribution is univariate, we will usually be able to generate random draws from it by making use of one of the algorithms for non-uniform random variate generation found in Devroye [7]. A number of these algorithms are also included in the syllabus for Associateship Examination Part 4B (e.g., Hogg and Klugman [16, pp. 69–75]). Gilks [12], Gilks and Wild [13], and Wild and Gilks [32] describe clever adaptive rejection sampling (ARS) methods that are very efficient when random draws are required from a univariate continuous distribution with a density that is concave with respect to its argument on the logarithmic scale, and these methods are becoming very popular as well. Many of these algorithms, including those for ARS in particular, do not even necessitate the calculation of the normalizing constants.

If one decides to implement a Gibbs sampler by coding it directly using a high-level programming language like APL, C, or FORTRAN, it will probably be necessary to code one or more of the algorithms for non-uniform random variate generation mentioned above. One way to avoid this bother is to make use of an existing software package for statistical computing, like S-Plus (Statistical Sciences Inc.) (Becker, Chambers, Wilks [1]) or Minitab (Minitab Inc.). Using a statistical computing package is often a convenient way in which to implement a Gibbs sampler, since random number generators for many standard distributions are often included in these packages. We implemented the Gibbs samplers for Examples 1 and 2 within the S-Plus programming environment using the random number generators rgamma, rbinom, and rpois, and each of the simulations took only seconds to run. On the other hand, intensive MCMC simulations for more complicated models often take minutes or hours to run when implemented using Minitab or S-Plus, but require only a few seconds or minutes to run when programmed in a high-level language like C or FORTRAN.

Specialized software for Gibbs sampling also exists. Foremost is the software package known as BUGS (Thomas, Spiegelhalter, and Gilks [30] and Gilks, Thomas, and Spiegelhalter [14]). Its name is an acronym for *Bayesian Inference Using Gibbs Sampling*, and BUGS is intended to be used for that purpose. BUGS will implement Bayesian inference using Gibbs sampling for a large class of full probability models in which all quantities are treated as random variables. This package is capable of analyzing very complicated models, and it appears to be competitive with C or FORTRAN in terms of raw speed. The BUGS software package is very convenient to use, insomuch as it provides a declarative language permitting the practitioner to make a straightforward specification of the statistical model at hand, following which the software automatically derives the associated full conditional distributions and selects appropriate sampling algorithms. Version 0.50 of this software is available free of charge for SUN Sparcstations and PC 386+387/486/586 platforms. Readers with access to the computer Internet may obtain BUGS, along with an instruction manual (Spiegelhalter, Thomas, Best, and Gilks [27]) and two volumes of worked examples (Spiegelhalter, Thomas, Best, and Gilks [26]) by anonymous ftp from *ftp.mrc-bsu.cam.ac.uk* in the directory *pub/methodology/bugs* or by accessing the uniform resource locator *http://www.mrc-bsu.cam.ac.uk* on the World Wide Web. These resources may also be obtained on disk from the developers for a small administrative fee. (For details, e-mail *bugs@mrc-bsu.cam.ac.uk*.) In Appendices A, B, and C, we provide illustrative BUGS programming code corresponding to Examples 3 and 4, which are themselves presented in Section 5. After reviewing Example 4, the reader will recognize that BUGS requires relatively few lines of code in order to implement a Gibbs sampler, even for a large and complicated model.

Recall that it will be necessary to run a Gibbs sampler for a little while in order to escape from the influence of the initial values and converge to the target distribution. Regardless of how one chooses to implement a Gibbs sampler, it will always be necessary to monitor this convergence. This is usually best diagnosed on the basis of the output from several independent replications of the Gibbs sampler, using widely dispersed starting values. If these Gibbs samplers have been left to run for a sufficiently long time so that convergence has been obtained, then the inferences drawn from each of the replications should be consistent, and virtually identical, with one another. In a similar vein, the behavior of the sampled values across replications at various iterations should be consistent when a large number of replications is considered. An ad hoc implementation of this idea is used in Example 3. More formal diagnostics are also available, and Cowles and Carlin [6] recently made a comparative review of a number of these. Two of the most popular are the methods proposed by Gelman and Rubin [10] and Raftery and Lewis [22]. An application of Gelman and Rubin's method may be found in Scollnik [24].

## 5. BAYESIAN ANALYSIS USING GIBBS SAMPLING

The Gibbs sampler has proven itself to be particularly suited for problems arising in the field of Bayesian statistical inference. Recall that a Bayesian analysis proceeds by assuming a model $f(Y \mid \theta)$ for the data $Y$ conditional upon the unknown parameters $\theta$. When $f(Y \mid \theta)$ is considered as a function of $\theta$ for fixed $Y$, it is referred to as the likelihood and is denoted by $L(\theta \mid Y)$ or $L(\theta)$. A prior probability distribution $f(\theta)$ describes our knowledge of the model parameters before the data is actually observed. Bayes' theorem allows us to combine the likelihood function with the prior in order to form the conditional distribution of $\theta$ given the observed data $Y$, that is,

$$f(\theta \mid Y) \propto f(\theta)L(\theta). \tag{5.1}$$

This conditional distribution is called the posterior distribution for the model parameters, and describes our updated knowledge of them after the data has been observed. Frequently, numerical methods are required in order to study posterior distributions with complicated forms. Following Equation 4.1 and its associated discussion, one may deduce that the Gibbs sampler is one method available for consideration. Other numerical methods that might be utilized in order to advance a Bayesian analysis include numerical quadrature and Monte Carlo integration, both of which are described in Klugman [18, Chapter 2]. One of the big advantages of the Gibbs sampler is that it is often far easier to implement than either of these other two methods. The Gibbs sampler is also flexible in the sense that its output may be used in order to make a variety of posterior and predictive inferences.

For example, imagine that we have implemented a Gibbs sampler generating values $\theta^{(i)}$ from $f(\theta \mid Y)$, provided that $i$ is sufficiently large. Obviously, posterior inference with respect to $\theta$ may proceed on the basis of the sampled values $\theta^{(i)}$. However, if some transformation $\omega = \omega(\theta)$ of the model parameters is of interest as well, then posterior inference with respect to $\omega$ is immediately available on the basis of the transformed values $\omega^{(i)} = \omega(\theta^{(i)})$. Further, it will often be the case that the actuarial practitioner will be interested in making predictive inferences with respect to things like future claim frequencies, future size of losses, and so forth. Typically, the conditional model $f(Y_f \mid Y,\theta)$ for the future data $Y_f$ given the past data $Y$ and the model parameters $\theta$ will be available. The appropriate distribution upon which to base future inferences is the so-called predictive distribution with density

$$f(Y_f \mid Y) = \int f(Y_f \mid Y,\theta) f(\theta \mid Y)\, d\theta, \tag{5.2}$$

which describes our probabilistic knowledge of the future data given the observed data. An estimate of this predictive density is easily obtained by averaging $f(Y_f \mid Y,\theta)$ over the sampled values of $\theta^{(i)}$ in the sense of Equation 2.2. Recall that the density estimates appearing in Equations 1.3, 1.4, and 3.7 were all constructed in a like manner.

This section concludes with Examples 3 and 4. Example 3 involves the estimation of the parameter for a size of loss distribution when grouped data are observed. Example 4 addresses credibility for classification ratemaking via hierarchical models, and involves the prediction of frequency counts in workers compensation insurance. We will operate within the Bayesian paradigm for these examples and implement the Bayesian analyses using the Gibbs sampler.

*Example 3*

Assume that loss data has been generated according to the Pareto$(\theta, \lambda)$ distribution with density

$$f(x \mid \theta) = \frac{\theta\lambda^{\theta}}{(\lambda + x)^{\theta+1}}, \qquad 0 < x < \infty. \tag{5.3}$$

In order to simplify the presentation, we will assume that the parameter $\lambda$ is known to be equal to 5000, so that the only uncertainty is with respect to the value of the parameter $\theta$. Imagine that twenty-five independent observations are available in total, but that the data has been grouped in such a way so that we know only the class frequencies: 12, 8, 3, and 2 observations fall into the classes (0,1000], (1000,2000], (2000,3000], (3000,$\infty$), respectively. Hogg and Klugman [16, pp. 81–84] consider maximum likelihood, minimum distance, and minimum chi-square estimation for grouped data problems like this when inference is sought with respect to the parameter $\theta$. Below, we will consider how a Bayesian analysis might proceed.

Given the situation described in the paragraph above, the best likelihood function available is proportional to

$$L(\theta \mid \mathit{Obs.\ Data}) = \prod_{i=1}^{4} \left( \int_{c_{i-1}}^{c_i} f(x \mid \theta)\, dx \right)^{f_i}$$

with class limits $c_0 = 0$, $c_1 = 1000$, $c_2 = 2000$, $c_3 = 3000$, $c_4 = \infty$, and class frequencies $f_1 = 12$, $f_2 = 8$, $f_3 = 3$, $f_4 = 2$. Multiplying this likelihood function together with a prior density for $\theta$ will result in an expression proportional to the posterior density for $\theta$ given the observed data. Since the posterior distribution of $\theta$ is univariate, this posterior density might be evaluated in a straightforward fashion making use of numerical quadrature methods.

However, for the sake of illustration, we choose instead to implement the Bayesian analysis by utilizing the Gibbs sampler along with a process called data augmentation. Towards this end, let us first consider how the likelihood function would change if exact size of loss values supplementing or augmenting the twenty-five observed class frequencies were available as well. In this case, the likelihood function would be proportional to

$$L(\theta \mid \mathit{Obs.\ \&\ Aug.\ Data}) = \frac{\theta^{25}\lambda^{25\theta}}{\prod_{i=1}^{25}(\lambda + x_i)^{\theta+1}}.$$

Combining this likelihood function with the conjugate gamma($\alpha,\beta$) prior density for $\theta$ results in the posterior density

$$\begin{aligned} &f(\theta \mid \mathit{Obs.\ \&\ Aug.\ Data}) \\ &\quad \propto \theta^{24+\alpha}\exp\left(-\theta\left(\beta - 25\ln\lambda + \sum_{i=1}^{25}\ln\,[\lambda + x_i]\right)\right) \\ &\quad \sim \text{gamma}\left(25+\alpha, \beta - 25\ln\lambda + \sum_{i=1}^{25}\ln\,[\lambda + x_i]\right). \qquad (5.4) \end{aligned}$$

Recall that a conjugate prior combines with the likelihood function in such a way so that the posterior distribution has the same form as the prior. For this example, we adopted the conjugate prior primarily for mathematical and expository convenience, and set $\alpha = \beta = 0.001$ so that our prior density for $\theta$ is very diffuse and noninformative with mean 1 and variance 1000. Although the the adoption of a diffuse conjugate prior is not uncommon when relatively little prior information is being assumed, in practice the practitioner should adopt whatever form of prior density that best describes the prior information actually available.

Now, the augmented data values are all independently distributed given the model parameters, and each is distributed according to Equation 5.3 but restricted to the appropriate class interval. In other words, the conditional distribution of the augmented data, given the model parameters and the observed class frequencies, is described by the following set of truncated Pareto distributions:

$$x_i \sim \text{truncated Pareto } (\theta, \lambda) \text{ on the interval } (0, 1000], \quad \text{for} \quad i = 1, 2, \ldots, 12; \tag{5.5}$$

$$x_i \sim \text{truncated Pareto } (\theta, \lambda) \text{ on the interval } (1000, 2000], \quad \text{for} \quad i = 13, 14, \ldots, 20; \tag{5.6}$$

$$x_i \sim \text{truncated Pareto } (\theta, \lambda) \text{ on the interval } (2000, 3000], \quad \text{for} \quad i = 21, 22, 23; \tag{5.7}$$

$$x_i \sim \text{truncated Pareto } (\theta, \lambda) \text{ on the interval } (3000, \infty), \quad \text{for} \quad i = 24, 25. \tag{5.8}$$

If a loss random variable has a Pareto distribution, with parameters $\theta$ and $\lambda$ and a density function as in Equation 5.3, then the truncated density function of that random variable, on the restricted interval $(l, u]$, with $0 \leq l < u \leq \infty$, is given by

$$\frac{f(x \mid \theta)}{\Pr(l < X \leq u \mid \theta)}, \qquad l < x \leq u.$$

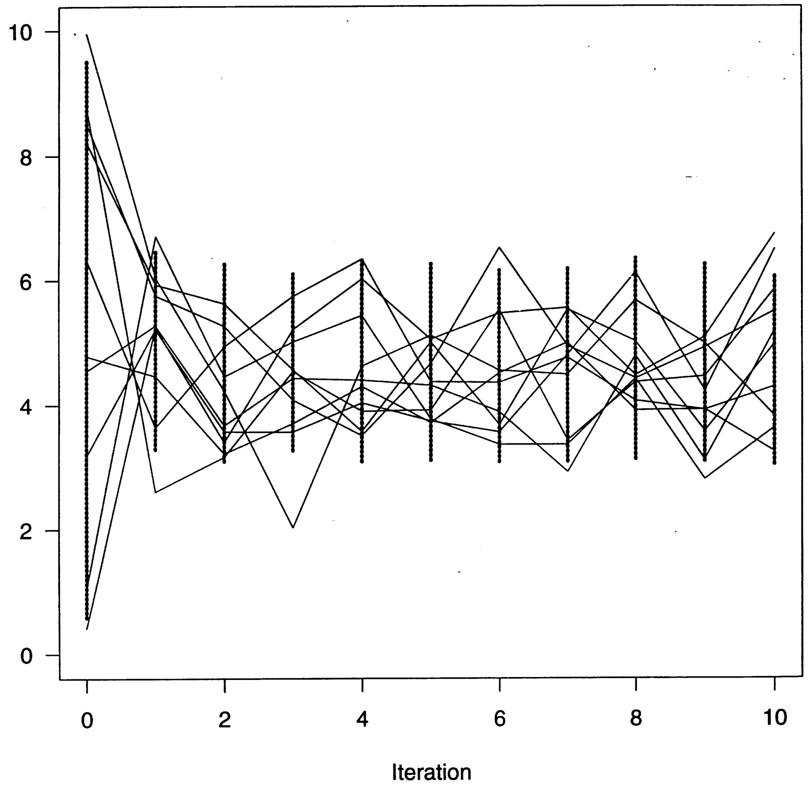
For example, the density function associated with the truncated Pareto distribution appearing in Equation 5.8 is given by

$$\frac{\theta(\lambda + 3000)^{\theta}}{(\lambda + x)^{\theta+1}}, \qquad 3000 < x < \infty.$$

By applying the Gibbs sampler to the 26 full conditional distributions defined by Equations 5.4 through 5.8, we are easily able to simulate a Markov chain with an invariant distribution equal to $p(\theta, \mathit{Aug.\ Data} \mid \mathit{Obs.\ Data})$. In order to make posterior inference with respect to $\theta$, the parameter of interest, we initiated

FIGURE 6

TEN SAMPLE PATHS FOR $\theta$ IN EXAMPLE 3



1000 replications of this Markov chain, using randomly selected starting values for $\theta$ and the augmented data each time, and let each replication run for 10 iterations. Only the values generated in the final iteration of each replication will be used. The reader will recall that $\lambda$ is equal to 5000 by assumption.

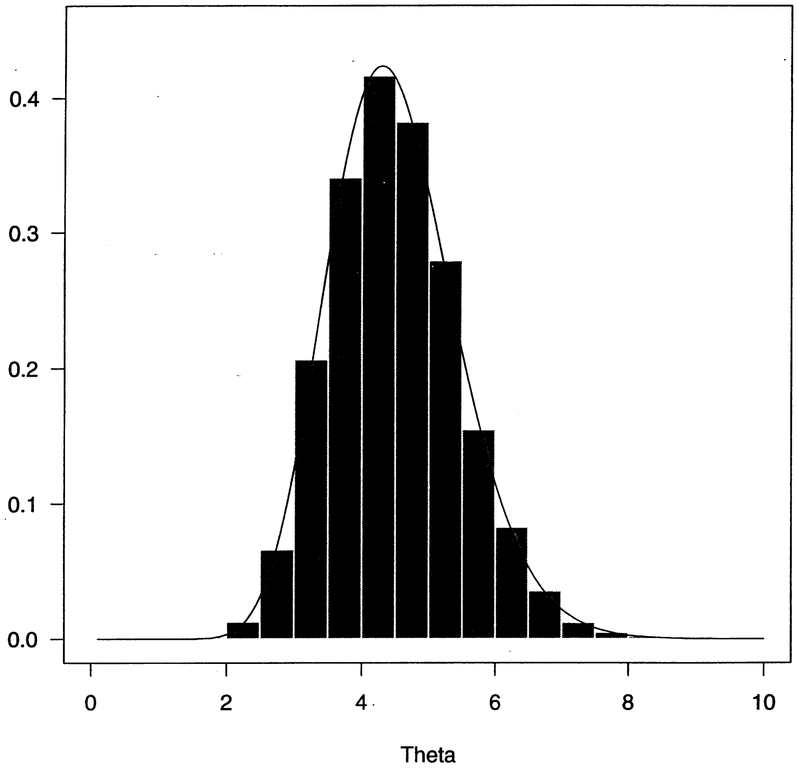
Ten arbitrarily selected sample paths for $\theta$ are plotted in Figure 6 for illustrative purposes. These 10 sample paths are typical of the entire collection of 1000 generated sample paths, and indicate that the simulated sequences stabilize almost immediately. In order to monitor convergence, we monitored the 5th and 95th empirical quantiles for $\theta$ over the 1000 replications for each iteration. The resulting 90% empirical confidence bands are plotted as vertical lines in Figure 6. These stabilize almost immediately as well, indicating that the Gibbs sampler was very quick to converge. Although we do not include the plots, the sample paths for the augmented data behaved similarly. By taking only the final value of $\theta$ appearing in each of the 1000 sample paths, we obtain an approximately independent random sample from the posterior distribution $p(\theta \mid \textit{Obs. Data})$. These values were used to construct the histogram of sampled values for $\theta$ appearing in Figure 7. Their sample mean and variance were 4.5097 and 0.9203, respectively. A smooth density estimate for $\theta$ was obtained by averaging Equation 5.4 over the corresponding 1000 sets of simulated values for the augmented data, and this estimate overlays the histogram in Figure 7. At this time we note that the data yielding the observed class frequencies used in this example were actually generated using a value of $\theta$ equal to 4.5, so that our posterior inference with respect to $\theta$ is certainly on the mark.

By monitoring the values taken on by the augmented data as the simulation proceeds, we can also make posterior inference with respect to the actual but unobserved sizes of loss. For example, by monitoring the values of the two losses appearing in the upper-most class, $x_{24}$ and $x_{25}$, we can estimate the posterior probability that one or more of these two losses exceeded an upper limit of, say, 10,000 by simply observing the proportion of times this event occurred in the simulation. In fact, we observed that in only 124 of the 1000 final pairs of simulated values for $x_{24}$ and $x_{25}$ did at least one of these two values exceed 10,000. Thus, a simple estimate of the posterior probability of interest is given by the binomial proportion

$$\hat{p} = \frac{124}{1000} = 0.124,$$

FIGURE 7

HISTOGRAM OF SAMPLED VALUES AND A DENSITY ESTIMATE FOR $\theta$ IN EXAMPLE 3



which has an estimated standard error of

$$\left(\frac{\hat{p}(1-\hat{p})}{1000}\right)^{0.5} = \left(\frac{0.124 * 0.876}{1000}\right)^{0.5} = 0.0104.$$

The analysis described above was implemented using the statistical computing package S-Plus on a SUN Sparcstation LX (operating at 50 MHz). Five thousand iterations of the Gibbs sampler constructed for this problem took 380 seconds. By way of comparison, when implemented using BUGS on the same computer, 5000 iterations of the Gibbs sampler took 14 seconds. These times should be comparable to those one might encounter using a fast 486/586 PC. The BUGS code corresponding to this example appears in Appendix A.

Although we assumed that $\lambda$ was known to be equal to 5000 in this example, treating it as a random parameter would have complicated the analysis only slightly. In this case, it would have been necessary to include random draws from the posterior full conditional distribution for $\lambda$, given $\theta$ and the augmented data, in the running of the Gibbs sampler. These univariate random draws might have been accomplished utilizing one of the strategies for random number generation described in Section 4. Of course, a prior distribution for the parameter $\lambda$ would have to have been specified as well.

*Example 4*

For our last example, we will use the Gibbs sampler to analyze three models for claim frequency counts. These are the hierarchical normal, hierarchical first level Poisson, and the variance stabilized hierarchical normal models. The data corresponds to Data Set 2 in Klugman [18]. The observations are frequency counts in workers compensation insurance. The data were collected from 133 occupation classes over a seven-year period. The exposures are scaled payroll totals adjusted for inflation. The first two classes are given in Table 1, and the full data set may be found in Appendix F of Klugman [18]. Only the first six of the seven years will be used to analyze these models, and omitting those cases with zero exposure yields a total of 767 observations. The results of each model analysis will then be used to forecast the number of claims associated with the 128 classes with non-zero exposure in the seventh year. We will observe that the second of the three models (i.e., the hierarchical first level Poisson model) appears to have associated with it the best predictive performance in this context.

TABLE 1

WORKERS COMPENSATION INSURANCE FREQUENCIES

| Class | Year | Exposure | Claims |
|---|---|---|---|
| $i$ | $j$ | $P_{ij}$ | $Y_{ij}$ |
| 1 | 1 | 32.322 | 1 |
| 1 | 2 | 33.779 | 4 |
| 1 | 3 | 43.548 | 3 |
| 1 | 4 | 46.686 | 5 |
| 1 | 5 | 34.713 | 1 |
| 1 | 6 | 32.857 | 3 |
| 1 | 7 | 36.600 | 4 |
| 2 | 1 | 45.995 | 3 |
| 2 | 2 | 37.888 | 1 |
| 2 | 3 | 34.581 | 0 |
| 2 | 4 | 28.298 | 0 |
| 2 | 5 | 45.265 | 2 |
| 2 | 6 | 39.945 | 0 |
| 2 | 7 | 39.322 | 4 |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |

Klugman [17] argued that a hierarchical model is the most appropriate framework in which to implement credibility for classification ratemaking. In this spirit, Klugman [18] considered a Bayesian analysis of the workers compensation insurance frequency count data presently under consideration using a (one-way) hierarchical normal model (HNM), and demonstrated that this analysis might be implemented using any one of a number of numerical techniques, emphasizing numerical quadrature, Monte Carlo integration, or Tierney–Kadane's integral method. We will begin by considering the HNM as well, but we will implement its analysis using the Gibbs sampler. Letting $x_{ij}$ denote $Y_{ij}/P_{ij}$, the relative frequency for class $i$ and year $j$, the first two levels of the HNM we consider are described by

$$f(x_{ij} \mid \theta_i, \tau_1^2) \sim \text{normal } (\theta_i, P_{ij}\tau_1^2) \tag{5.9}$$

$$\text{and} \qquad f(\theta_i \mid \mu, \tau_2^2) \sim \text{normal } (\mu, \tau_2^2). \tag{5.10}$$

Each of these normal densities is indexed by two parameters, a mean and a precision (i.e., inverse variance). The model parameter $\theta_i$ represents the true relative frequency for the $i$th class. The relative frequencies $x_{ij}$, for $i = 1,\ldots,133$ and $j = 1,\ldots,7$, are assumed to be independent across class and year, given the underlying model parameters $\theta_i$, $i = 1,\ldots,133$, and $\tau_1^2$. Similarly, the true frequencies $\theta_i$, for $i = 1,\ldots,133$, are assumed to be independent, given the underlying parameters $\mu$ and $\tau_2^2$. (Notice that under this model, negative claim frequencies are possible. For this reason, the HNM as presented is not entirely appropriate for modeling the non-negative workers compensation insurance frequency count data. We return to this point in the next paragraph.) In order to complete the model specification, Klugman [18] employed a constant improper prior density for the model parameters $\mu$, $\tau_1^2$ and $\tau_2^2$. Instead, we adopt the diffuse but proper prior density described by

$$f(\mu) \sim \text{normal } (0, 0.001), \tag{5.11}$$

$$f(\tau_1^2) \sim \text{gamma } (0.001, 0.001), \tag{5.12}$$

$$f(\tau_2^2) \sim \text{gamma } (0.001, 0.001). \tag{5.13}$$

The assumption of a proper prior guarantees that the posterior distribution exists and is proper as well, and slightly simplifies the implementation of a Gibbs sampler. However, the precise form of the diffuse prior (and the selection of the prior density parameters) is not terribly important in this instance since the observed data comprises a rather large sample that will tend to dominate the prior information in any case. (Also, see the discussion in the next paragraph.) We assume prior independence between the model parameters $\mu$, $\tau_1^2$ and $\tau_2^2$. If we assume that only the data corresponding to the first six of the seven years has been observed, then the posterior density for the model parameters is proportional to the product of terms appearing on the right-hand side of the expression

$$f(\theta_1,\theta_2,\ldots,\theta_{133},\mu,\tau_1^2,\tau_2^2 \mid \mathit{Data}) \propto f(\mu)f(\tau_1^2)f(\tau_2^2)\prod_{i=1}^{133} f(\theta_i \mid \mu,\tau_2^2)\prod_{j=1}^{6} f(x_{ij} \mid \theta_i,\tau_1^2). \tag{5.14}$$

Recall that our objective is to forecast the number of claims associated with the 128 classes with non-zero exposure in the seventh year.

As previously remarked, the HNM as presented above is not entirely appropriate for modeling the non-negative workers compensation insurance frequency count data. Yet, we will continue with its analysis for the following reasons:

- the large amount of sample data available will tend to overwhelm the prior density and will also go a long way towards correcting the inadequacy of the model by assigning less posterior probability to parameter values that are likely to generate negative frequencies;
- it will be interesting to compare the results of the MCMC simulation based analysis of the HNM to the numerical analysis presented by Klugman [18];
- the MCMC simulation based analysis of the HNM provides a benchmark to which the MCMC simulation based analyses of the other two models may be compared; and
- the HNM is a very important model in its own right, and for this reason alone it is valuable and instructive to see how its Gibbs sampling based Bayesian analysis might proceed.

Having said this, there are at least two ways in which the basic HNM may be constructively modified if it is to be applied to frequency count data. The first solution is to adopt the recommendation made by Klugman [18, pp. 76–77] and transform the original data in some way so that the transformed data is more appropriately modeled using the HNM. This approach motivates the variance stabilized hierarchical normal model considered at the end of this section. The second solution involves adding restrictions to the HNM so that negative frequencies are prohibited. A simple way in which to do this is to replace the normal distributions appearing in Equations 5.9, 5.10, and 5.11 with normal distributions truncated below at zero. In fact, we analyzed this truncated HNM and observed that, although it did perform significantly better than the non-truncated HNM, it did not perform as well as the variance stabilized hierarchical normal model. For this reason, we will omit the details of the truncated HNM analysis.

In order to conduct a Bayesian analysis of the HNM described by Equations 5.9 to 5.13 using the Gibbs sampler, we are required to first derive the necessary full conditional distributions associated with this model. These may be derived by substituting Equations 5.9 through 5.13 into Equation 5.14, and then making use of the discussion following Equation 4.1. In this manner, for an arbitrary one of the 133 normal mean $\theta_i$ parameters we obtain

$$
\begin{aligned}
&f(\theta_i \mid \mathit{Data};\ \theta_1,\ldots,\theta_{i-1},\theta_{i+1},\ldots,\theta_{133},\mu,\tau_1^2,\tau_2^2) \\
&\quad \propto f(\theta_i \mid \mu,\tau_2^2)\prod_{j=1}^{6} f(x_{ij} \mid \theta_i,\tau_1^2) \\
&\quad \propto \exp\left(-0.5\left[\tau_2^2(\theta_i-\mu)^2+\tau_1^2\sum_{j=1}^{6}P_{ij}(x_{ij}-\theta_i)^2\right]\right) \\
&\quad \sim \text{normal}\left(\left[\mu\tau_2^2+\tau_1^2\sum_{j=1}^{6}P_{ij}x_{ij}\right]\Bigg/\right. \\
&\qquad\qquad\qquad \left.\left[\tau_2^2+\tau_1^2\sum_{j=1}^{6}P_{ij}\right],\tau_2^2+\tau_1^2\sum_{j=1}^{6}P_{ij}\right).
\end{aligned}
$$

The recognition of the normal distribution in the last line of this derivation follows from completing the square in $\theta_i$ in the expression immediately above it. Observe that the full conditional distribution of $\theta_i$ is actually independent of any other parameter $\theta_j$. For the parameter $\mu$, we obtain

$$\begin{aligned}
&f(\mu \mid \mathit{Data};\ \theta_1,\ldots,\theta_{133},\tau_1^2,\tau_2^2)\\
&\qquad \propto f(\mu)\prod_{i=1}^{133} f(\theta_i \mid \mu,\tau_2^2)\\
&\qquad \propto \exp\left(-0.5\left[0.001\mu^2 + \tau_2^2\sum_{i=1}^{133}(\theta_i-\mu)^2\right]\right)\\
&\qquad \sim \text{normal}\left(\left[\tau_2^2\sum_{i=1}^{133}\theta_i\right]\Big/[0.001 + 133\tau_2^2],\ 0.001 + 133\tau_2^2\right);
\end{aligned}$$

and for the precision parameter $\tau_1^2$, we have

$$\begin{aligned}
&f(\tau_1^2 \mid \mathit{Data};\ \theta_1,\ldots,\theta_{133},\mu,\tau_2^2)\\
&\qquad \propto f(\tau_1^2)\prod_{i=1}^{133}\prod_{j=1}^{6} f(x_{ij} \mid \theta_i,\tau_1^2)\\
&\qquad \propto (\tau_1^2)^{0.001-1+767/2}\\
&\qquad\quad \times \exp\left(-\tau_1^2\left[0.001 + 0.5\sum_{i=1}^{133}\sum_{j=1}^{6}P_{ij}(x_{ij}-\theta_i)^2\right]\right)\\
&\qquad \sim \text{gamma}\left(383.501, 0.001 + 0.5\sum_{i=1}^{133}\sum_{j=1}^{6}P_{ij}(x_{ij}-\theta_i)^2\right).
\end{aligned}$$

In the derivation of the full conditional distribution for $\tau_1^2$ we made use of the fact that there were only 767 observations associated with non-zero exposures in the first six years. Finally, for the precision parameter $\tau_2^2$, we have

$$
\begin{aligned}
&f(\tau_2^2 \mid Data;\ \theta_1,\ldots,\theta_{133},\mu,\tau_1^2)\\
&\qquad \propto f(\tau_2^2)\prod_{i=1}^{133} f(\theta_i \mid \mu,\tau_2^2)\\
&\qquad \propto (\tau_2^2)^{0.001-1+133/2}\exp\left(-\tau_2^2\left[0.001+0.5\sum_{i=1}^{133}(\theta_i-\mu)^2\right]\right)\\
&\qquad \sim \text{gamma}\left(66.501, 0.001+0.5\sum_{i=1}^{133}(\theta_i-\mu)^2\right).
\end{aligned}
$$

Using these full conditional distributions, we can implement a Gibbs sampler for the model of interest by coding the appropriate random number generators in a high-level programming language like APL, C, or FORTRAN. However, for this moderately large model incorporating 136 random parameters, we prefer to use the BUGS software package mentioned in Section 4 and allow it to automatically select and program the necessary random number generators for us. In fact, since BUGS automatically determines and selects the appropriate random number generators for the full conditional distributions directly from Equations 5.9 through 5.13, we really did not have to derive these full conditional distributions ourselves, except perhaps to demonstrate how this task is accomplished.

Illustrative BUGS code corresponding to the HNM of interest is provided in Appendix B, and we used this programming code in conjunction with the BUGS software package in order to implement a Gibbs sampler for the problem at hand. We allowed the MCMC simulation to run for 20,000 iterations in order to "burn-in" the Gibbs sampler and remove the effect of the starting values, and then allowed it to run for an additional 5000 iterations in order to generate a dependent random sample of size

5000 from the joint posterior distribution of the model parameters $\theta_1, \theta_2, \ldots, \theta_{133}$, $\mu$, $\tau_1^2$, and $\tau_2^2$, given the observed frequency counts. (This simulation took about 337 seconds on the same SUN Sparcstation LX we described previously.)

The values generated by this MCMC simulation may be used in order to make a wide variety of posterior and predictive inferences. For example, the expected number of claims for year seven in class $i$ is given by $\mathrm{E}(Y_{i7} \mid P_{i7}, \theta_i) = P_{i7}\theta_i$. By multiplying each of the 5000 simulated values of $\theta_i$ with $P_{i7}$, we obtain 5000 realizations from the posterior distribution of $P_{i7}\theta_i$. Then posterior inference with respect to the expected number of claims for year seven in class $i$ may proceed on the basis of this sample. We performed this procedure for five of the 128 classes with non-zero exposure in the seventh year, and have recorded in Table 2 the empirical mean, standard deviation, and several quantiles (i.e., the 2.5th, 50.0th, and 97.5th) for each of the resulting samples of size 5000. The five classes we selected are the same (non-degenerate) ones considered by Klugman [18, p. 128]. However, whereas Klugman provided only point estimates for the expected number of claims for year seven in each of these classes, we have been able to generate realizations from the posterior distribution of these expected claim numbers using MCMC. This allows us to observe, for instance, that both classes 70 and 112 have substantial posterior probability associated with negative expected number of claim values under the HNM, as is evidenced by Table 2. An overall measure of this model's prediction success is given by the statistic

$$OMPS = \sum_{P_{i7}>0} \frac{(P_{i7}\theta_i - Y_{i7})^2}{P_{i7}}. \tag{5.15}$$

(The name of this statistic is an abbreviation of Overall Measure of Prediction Success.) There are 128 terms in the summation, one for each of the 128 classes with non-zero exposure in the seventh year. Small values of *OMPS* are indicative of a model with good overall prediction success. We obtained 5000

## TABLE 2

### EXPECTED WORKERS COMPENSATION INSURANCE FREQUENCIES (HIERARCHICAL NORMAL MODEL)

| | Actual Values | | Expected Number of Claims $P_{i7}\theta_i$ | | | | |
|---|---|---|---|---|---|---|---|
| Class | Exposure | Claims | Estimated Posterior Summary Statistics | | | | |
| $i$ | $P_{i7}$ | $Y_{i7}$ | Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 11 | 229.83 | 8 | 10.19 | 2.98 | 4.34 | 10.21 | 16.04 |
| 20 | 1,315.37 | 22 | 41.38 | 5.57 | 30.62 | 41.38 | 52.30 |
| 70 | 54.81 | 0 | 0.61 | 1.20 | −1.74 | 0.61 | 2.95 |
| 89 | 79.63 | 40 | 29.42 | 1.47 | 26.54 | 29.44 | 32.27 |
| 112 | 18,809.67 | 45 | 36.11 | 27.55 | −19.02 | 36.47 | 89.43 |

realizations from the posterior distribution of *OMPS* by simply evaluating Equation 5.15 five thousand times, once using each of the 5000 joint realizations of $\theta_1, \theta_2, \ldots, \theta_{133}$ previously simulated from their posterior joint distribution. In Table 3, we present the empirical mean, standard deviation, and 2.5th, 50th and 97.5th quantiles for this sample of 5000 realizations from the posterior distribution of *OMPS*. We will return to this table after we introduce and analyze our second model. Incidentally, we remark that we checked our inferences throughout this example by independently replicating our entire MCMC simulation-based analysis several times, using different starting values for the Gibbs sampler each time.

Above, we concentrated on posterior inferences made with respect to the expected number of claims in year seven for various classes. As remarked at the start of Section 5, if we are interested in the future number of actual claims, then we should really be using the relevant predictive distribution in order to fashion our inferences. For the HNM presently under consideration, the distribution of the future number of claims for year seven in class $i$ is independent of the data associated with the first six years provided that the underlying model parameters are

TABLE 3

AN OVERALL MEASURE OF PREDICTION SUCCESS (HIERARCHICAL NORMAL MODEL)

| Estimated Posterior Summary Statistics | | | | |
|---|---|---|---|---|
| Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 16.49 | 1.23 | 14.16 | 16.48 | 18.96 |

known. This distribution is given by

$$f(Y_{i7} \mid \theta_i, \tau_1^2) \sim \text{normal}\left(P_{i7}\theta_i, \frac{\tau_1^2}{P_{i7}}\right), \tag{5.16}$$

in accord with Equation 5.9. Following Equation 5.2, it follows that the predictive distribution of the future number of claims for year seven in class $i$ given the observed frequency counts over the first six years is

$$f(Y_{i7} \mid Data) = \int f(Y_{i7} \mid \theta_i, \tau_1^2) f(\theta_i, \tau_1^2 \mid Data)\, d\theta_i\, d\tau_1^2. \tag{5.17}$$

An estimate of this predictive distribution is easily obtained by simply averaging the density found in Equation 5.16 over the 5000 pairs of realized values for $\theta_i$ and $\tau_1^2$ previously simulated from the posterior distribution of the model parameters.

Another way in which to proceed is by generating 5000 realizations of $Y_{i7}$ according to the distribution in Equation 5.16, one realization per pair of values previously simulated for $\theta_i$ and $\tau_1^2$. Then the 5000 simulated values of $Y_{i7}$ represent a random sample from the predictive distribution in Equation 5.17, and the empirical distribution of this sample may be used to motivate predictive inference. Using this latter approach we simulated random samples of size 5000 from the predictive distribution in Equation 5.17 for each of the five classes we examined previously, and summary statistics for the samples from these predictive distributions appear in Table 4. From Table 4, we may

TABLE 4

PREDICTED WORKERS COMPENSATION INSURANCE FREQUENCIES
(HIERARCHICAL NORMAL MODEL)

| | Actual Values | | Predicted Number of Claims $Y_{i7}$ | | | | |
|---|---|---|---|---|---|---|---|
| Class | Exposure | Claims | Estimated Predictive Summary Statistics | | | | |
| $i$ | $P_{i7}$ | $Y_{i7}$ | Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 11 | 229.83 | 8 | 10.24 | 7.41 | −4.12 | 10.22 | 24.92 |
| 20 | 1,315.37 | 22 | 41.46 | 17.04 | 8.03 | 41.39 | 75.22 |
| 70 | 54.81 | 0 | 0.63 | 3.46 | −6.10 | 0.62 | 7.51 |
| 89 | 79.63 | 40 | 29.46 | 4.21 | 21.22 | 29.45 | 37.72 |
| 112 | 18,809.67 | 45 | 37.74 | 66.28 | −92.32 | 38.25 | 167.52 |

observe that the predictive distributions associated with future claim frequencies exhibit greater variability than do the corresponding posterior distributions associated with the expected numbers of future claims. Also notice that 3 of the 5 classes (i.e., classes 11, 70, and 112) have substantial predictive probability associated with negative number of claim values in year seven under the HNM.

The second model we consider is a more realistic one for modeling frequency count data. This model is also hierarchical, and its first two levels are described by

$$f(Y_{ij} \mid \theta_i) \sim \text{Poisson } (P_{ij}\theta_i) \tag{5.18}$$

$$\text{and} \qquad f(\ln\theta_i \mid \mu,\tau^2) \sim \text{normal } (\mu,\tau^2). \tag{5.19}$$

The model parameter $\theta_i$ now represents the true Poisson claim frequency rate for the $i$th class with one unit of exposure. The frequency counts $Y_{ij}$, for $i = 1,\ldots,133$ and $j = 1,\ldots,7$, are assumed to be independent across class and year, given the underlying model parameters $\theta_i$, $i = 1,\ldots,133$, and the Poisson claim frequency rate parameters $\theta_i$, for $i = 1,\ldots,133$, are assumed to be independent, given the underlying parameters $\mu$ and $\tau^2$. An obvious advantage of this model over the HNM considered previously is that the frequency counts are now being modeled at the first level with a discrete distribution on the non-negative integers. Assuming log-normal distributions as we have for the Poisson rate parameters implies that

$$\mathrm{E}(Y_{ij}) = \mathrm{E}(\mathrm{E}(Y_{ij} \mid \theta_i)) = P_{ij}\exp(\mu + 1/[2\tau^2]) = P_{ij}m \tag{5.20}$$

and

$$\begin{aligned}\mathrm{Var}(Y_{ij}) &= \mathrm{E}(\mathrm{Var}(Y_{ij} \mid \theta_i)) + \mathrm{Var}(\mathrm{E}(Y_{ij} \mid \theta_i)) \\ &= P_{ij}m + P_{ij}^2 m^2(\exp(1/\tau^2) - 1) > P_{ij}m, \end{aligned} \tag{5.21}$$

so that overdispersion is modeled in the count data. In order to complete the model specification, we will assume that the parameters $\mu$ and $\tau^2$ are independent a priori, and adopt the diffuse but proper prior density described by

$$f(\mu) \sim \text{normal } (0, 0.001), \tag{5.22}$$

$$f(\tau^2) \sim \text{gamma } (0.001, 0.001). \tag{5.23}$$

If we assume that only the data corresponding to the first six of the seven years has been observed, then the posterior density for the model parameters is proportional to the product of terms appearing on the right-hand side of the expression

$$\begin{aligned} &f(\theta_1, \theta_2, \ldots, \theta_{133}, \mu, \tau^2 \mid \mathit{Data}) \\ &\quad \propto f(\mu) f(\tau^2) \prod_{i=1}^{133} \theta_i^{-1} f(\ln\theta_i \mid \mu, \tau^2) \prod_{j=1}^{6} f(Y_{ij} \mid \theta_i). \end{aligned} \tag{5.24}$$

The $\theta_i^{-1}$ terms appearing in this expression arise from the change in variable when passing from $\ln\theta_i$ to $\theta_i$. As before, our objective is to forecast the number of claims associated with the 128 classes with non-zero exposure in the seventh year.

We will now conduct a Bayesian analysis of the (one-way) hierarchical model with a first level Poisson distribution, or hierarchical first level Poisson model (HFLPM), described above using the Gibbs sampler. Rather than derive the required full conditional distributions manually, and then program the necessary random number generators, we will let the BUGS software package do both tasks for us. Illustrative BUGS code corresponding to this model is provided in Appendix B, and we used this programming code in conjunction with the BUGS software package in order to implement a Gibbs sampler. As before, we allowed the MCMC simulation to run for 20,000 iterations in order to "burn-in" the Gibbs sampler and remove the effect of the starting values, and then allowed it to run for an additional 5000 iterations in order to generate a random sample of size 5000 from the joint posterior distribution of the model parameters $\theta_1, \theta_2, \ldots, \theta_{133}$, $\mu$, and $\tau^2$. This sample was used to implement the same sort of posterior and predictive inferences for the HFLPM as we did for the HNM considered previously.

We omit the specific details, but summaries of our estimated posterior and predictive inferences under the HFLPM are presented in Tables 5, 6, and 7. By comparing these summaries to those presented earlier in Tables 2, 3, and 4 for the HNM analysis, we are able to evaluate the relative performance of the two models. First of all, it is evident that the posterior and predictive distributions in which we are interested generally exhibit less variability under the HFLPM than under the HNM. Secondly, whereas the HNM permits negative relative frequencies and expected numbers of future claims, these are not a problem under the HFLPM. Finally, the posterior distribution of the *OMPS* statistic describing the overall measure of prediction success for a given model appears to be concentrated closer to zero under the HFLPM than under the HNM. In short, these observations suggest that the HFLPM may be a better model than the HNM for implementing credibility for classification ratemaking when the data is in terms of frequency counts.

Klugman [18, pp. 76–77, 152–153] also recognized that the HNM was inappropriate for modeling the workers compensation insurance frequency count data in its original form, and sug-

## TABLE 5

### EXPECTED WORKERS COMPENSATION INSURANCE FREQUENCIES (HIERARCHICAL FIRST LEVEL POISSON MODEL)

| | Actual Values | | Expected Number of Claims $P_{i7}\theta_i$ | | | | |
|---|---|---|---|---|---|---|---|
| Class | Exposure | Claims | Estimated Posterior Summary Statistics | | | | |
| $i$ | $P_{i7}$ | $Y_{i7}$ | Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 11 | 229.83 | 8 | 10.10 | 1.47 | 7.46 | 10.02 | 13.20 |
| 20 | 1,315.37 | 22 | 41.31 | 2.20 | 37.16 | 41.28 | 45.81 |
| 70 | 54.81 | 0 | 0.37 | 0.21 | 0.09 | 0.32 | 0.90 |
| 89 | 79.63 | 40 | 33.85 | 2.08 | 29.96 | 33.79 | 37.96 |
| 112 | 18,809.67 | 45 | 36.09 | 2.69 | 31.02 | 35.98 | 41.55 |

## TABLE 6

### AN OVERALL MEASURE OF PREDICTION SUCCESS (HIERARCHICAL FIRST LEVEL POISSON MODEL)

| Estimated Posterior Summary Statistics | | | | |
|---|---|---|---|---|
| Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 12.96 | 0.64 | 11.76 | 12.94 | 14.28 |

## TABLE 7

### PREDICTED WORKERS COMPENSATION INSURANCE FREQUENCIES (HIERARCHICAL FIRST LEVEL POISSON MODEL)

| | Actual Values | | Predicted Number of Claims $Y_{i7}$ | | | | |
|---|---|---|---|---|---|---|---|
| Class | Exposure | Claims | Estimated Predictive Summary Statistics | | | | |
| $i$ | $P_{i7}$ | $Y_{i7}$ | Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 11 | 229.83 | 8 | 10.12 | 3.48 | 4 | 10 | 18 |
| 20 | 1,315.37 | 22 | 41.36 | 6.82 | 28 | 41 | 55 |
| 70 | 54.81 | 0 | 0.38 | 0.65 | 0 | 0 | 2 |
| 89 | 79.63 | 40 | 33.76 | 6.19 | 22 | 34 | 47 |
| 112 | 18,809.67 | 45 | 36.18 | 6.57 | 24 | 36 | 49 |

gested that the variance stabilizing transformation $z_{ij} = 2\sqrt{x_{ij}}$ should first be applied to the relative frequencies $x_{ij} = Y_{ij}/P_{ij}$, for $i = 1,\ldots,133$ and $j = 1,\ldots,7$, in order to produce values that are approximately normal. The interested reader is referred to Klugman [18, p. 77] for a discussion of the rationale justifying this transformation. Klugman shows that if we apply this transformation and let $\gamma_i = 2\sqrt{\theta_i}$, then the appropriate variance stabilized hierarchical normal model (VSHNM) for the workers compensation data has its first two levels described by

$$f(z_{ij} \mid \gamma_i) \sim \text{normal } (\gamma_i, P_{ij}) \tag{5.25}$$

$$\text{and} \qquad f(\gamma_i \mid \mu, \tau^2) \sim \text{normal } (\mu, \tau^2). \tag{5.26}$$

As usual, these normal densities are indexed by two parameters, a mean and a precision (i.e., inverse variance). To complete the model specification, we adopt the diffuse proper priors

$$f(\mu) \sim \text{normal } (0, 0.001), \tag{5.27}$$

$$f(\tau^2) \sim \text{gamma } (0.001, 0.001), \tag{5.28}$$

and make our standard assumptions with respect to independence. We performed a Bayesian analysis of this model via the Gibbs sampler (using BUGS) and present summaries of the posterior and predictive analyses in Tables 8, 9, and 10. From Tables 3, 6, and 9, it appears that the VSHNM performed better than the original HNM, at least in terms of the posterior distribution of the statistic *OMPS* measuring overall prediction success, but not quite as well as the HFLPM. This observation is illustrated by Figure 8, in which we have plotted the estimated posterior distribution of *OMPS* resulting under each of the three hierarchical models.

## 6. CLOSING DISCUSSION

This paper focused on the MCMC method known as the Gibbs sampler. Other MCMC methods do exist. Perhaps the foremost of these is the so-called Metropolis-Hastings algorithm

FIGURE 8

ESTIMATED POSTERIOR DENSITIES FOR *OMPS* IN EXAMPLE 4

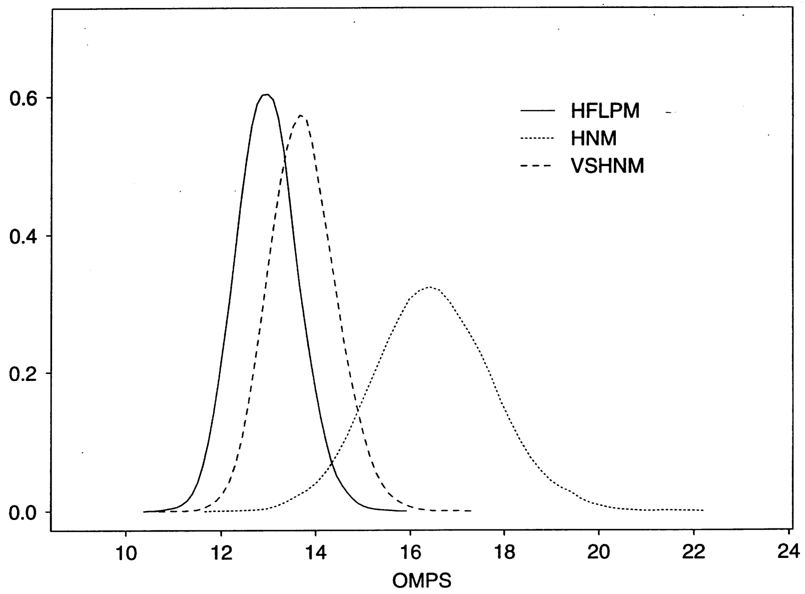


TABLE 8

EXPECTED WORKERS COMPENSATION INSURANCE FREQUENCIES (VARIANCE STABILIZED HIERARCHICAL NORMAL MODEL)

| | Actual Values | | Expected Number of Claims $P_{i7}\theta_i$ | | | | |
|---|---|---|---|---|---|---|---|
| Class | Exposure | Claims | Estimated Posterior Summary Statistics | | | | |
| $i$ | $P_{i7}$ | $Y_{i7}$ | Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 11 | 229.83 | 8 | 9.99 | 1.45 | 7.28 | 9.83 | 12.98 |
| 20 | 1,315.37 | 22 | 40.83 | 2.20 | 36.61 | 40.81 | 45.22 |
| 70 | 54.81 | 0 | 0.06 | 0.08 | 0 | 0.03 | 0.30 |
| 89 | 79.63 | 40 | 30.94 | 1.96 | 27.21 | 30.95 | 34.88 |
| 112 | 18,809.67 | 45 | 35.36 | 2.68 | 30.29 | 35.33 | 40.67 |

## TABLE 9

AN OVERALL MEASURE OF PREDICTION SUCCESS
(VARIANCE STABILIZED HIERARCHICAL NORMAL MODEL)

| Estimated Posterior Summary Statistics | | | | |
|---|---|---|---|---|
| Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 13.72 | 0.69 | 12.40 | 13.69 | 15.13 |

## TABLE 10

PREDICTED WORKERS COMPENSATION INSURANCE FREQUENCIES
(VARIANCE STABILIZED HIERARCHICAL NORMAL MODEL)

| | Actual Values | | Predicted Number of Claims $Y_{i7}$ | | | | |
|---|---|---|---|---|---|---|---|
| Class | Exposure | Claims | Estimated Predictive Summary Statistics | | | | |
| $i$ | $P_{i7}$ | $Y_{i7}$ | Mean | S.D. | 2.5% | 50.0% | 97.5% |
| 11 | 229.83 | 8 | 10.16 | 3.50 | 4.33 | 9.82 | 18.11 |
| 20 | 1,315.37 | 22 | 41.21 | 6.61 | 29.26 | 40.84 | 54.62 |
| 70 | 54.81 | 0 | 0.31 | 0.43 | 0 | 0.14 | 1.54 |
| 89 | 79.63 | 40 | 31.25 | 5.82 | 20.50 | 31.00 | 43.13 |
| 112 | 18,809.67 | 45 | 35.69 | 6.49 | 24.00 | 35.37 | 49.29 |

(Metropolis, Rosenbluth, Rosenbluth, Teller, Teller [20]; Hastings [15]; Roberts and Smith [23]). There are also many other actuarial problems beyond those discussed in this paper for which MCMC methods have potential application. These include the simulation of the aggregate claims distribution, the analysis of stochastic claims reserving models, and the analysis of credibility models with state-space formulations. We hope to report upon some of these topics and applications in the future.

## REFERENCES

[1] Becker, R. A., J. M. Chambers, and A. R. Wilks, *The New S Language*, Wadsworth & Brooks, Pacific Grove, California, 1988.

[2] Besag, J., "Spatial Interaction and the Statistical Analysis of Lattice Systems" (with discussion), *Journal of the Royal Statistical Society*, Series B, Vol. 36, 1974, pp. 192–326.

[3] Carlin, B. P., "State Space Modeling of Non-Standard Actuarial Time Series," *Insurance: Mathematics and Economics*, Vol. 11, 1992, pp. 209–222.

[4] Carlin, B. P., "A Simple Monte Carlo Approach to Bayesian Graduation," *Transactions of the Society of Actuaries* XLIV, 1992, pp. 55–76.

[5] Casella, G., and E. I. George, "Explaining the Gibbs Sampler," *The American Statistician*, Vol. 46, 1992, pp. 167–174.

[6] Cowles, M. K., and B. P. Carlin, "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," Research Report 94-008, Division of Biostatistics, School of Public Health, University of Minnesota, 1994. To appear in the *Journal of the American Statistical Association*.

[7] Devroye, L., *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

[8] Gelfand, A. E., "Gibbs Sampling," A Contribution to the *Encyclopedia of Statistical Sciences*, 1994.

[9] Gelfand, A. E., and A. F. M. Smith, "Sampling Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, Vol. 85, 1990, pp. 398–409.

[10] Gelman, A., and D. B. Rubin, "Inference from Iterative Simulation Using Multiple Sequences" (with discussion), *Statistical Science*, Vol. 7, No. 4, 1992, pp. 457–472.

[11] Geman, S., and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 6, 1984, pp. 721–741.

[12] Gilks, W. R., "Derivative-free Adaptive Rejection Sampling for Gibbs Sampling," *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., University Press, Oxford, 1992, pp. 641–649.

[13] Gilks, W. R., and P. Wild, "Adaptive Rejection Sampling for Gibbs Sampling," *Applied Statistics*, Vol. 41, No. 2, 1992, pp. 337–348.

[14] Gilks, W. R., A. Thomas, and D. J. Spiegelhalter, "A Language and Program for Complex Bayesian Modelling," *The Statistician*, Vol. 43, 1994, pp. 169–178.

[15] Hastings, W. K., "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, Vol. 57, 1970, pp. 97–109.

[16] Hogg, R. V., and S. A. Klugman, *Loss Distributions*, John Wiley & Sons, New York, 1984.

[17] Klugman, S. A., "Credibility for Classification Ratemaking via the Hierarchical Normal Linear Model," *PCAS* LXXIV, 1987, pp. 272–321.

[18] Klugman, S. A., *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*, Kluwer Academic Publishers, Norwell, 1992.

[19] Klugman, S. A., and B. P. Carlin, "Hierarchical Bayesian Whittaker Graduation," *Scandinavian Actuarial Journal*, Vol. 2, 1993, pp. 183–196.

[20] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics*, Vol. 21, 1953, pp. 1087–1092.

[21] Miller, R. B., and D. W. Wichern, *Intermediate Business Statistics: Analysis of Variance, Regression, and Time Series*, Holt, Rinehart and Winston, New York, 1977.

[22] Raftery, A. E., and S. Lewis, "How Many Iterations in the Gibbs Sampler?" in *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., University Press, Oxford, 1992, pp. 763–773.

[23] Roberts, G. O., and A. F. M. Smith, "Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms," *Stochastic Processes and their Applications*, Vol. 49, 1994, pp. 207–216.

[24] Scollnik, D. P. M., "A Bayesian Analysis of a Simultaneous Equations Model for Insurance Rate-Making," *Insurance: Mathematics and Economics*, Vol. 12, 1993, pp. 265–286.

[25] Smith, A. F. M., and G. O. Roberts, "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society*, Series B, Vol. 55, No. 1, 1993, pp. 3–23.

[26] Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks, *BUGS Examples 0.5*, Volumes 1 and 2, MRC Biostatistics Unit, Cambridge, 1995.

[27] Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks, *BUGS Manual 0.5*, MRC Biostatistics Unit, Cambridge, 1995.

[28] Tanner, M. A., and W. H. Wong, "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, Vol. 82, 1987, pp. 528–550.

[29] Tanner, M. A., *Tools for Statistical Inference, Methods for the Exploration of Posterior Distributions and Likelihood Functions*, second edition, Springer-Verlag, New York, 1993.

[30] Thomas, A., D. J. Spiegelhalter, and W. R. Gilks, "BUGS: A Program to Perform Bayesian Inference using Gibbs Sampling," *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., University Press, Oxford, 1992, pp. 837–842.

[31] Tierney, L., "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, Vol. 22, No. 4, 1994, pp. 1701–1728.

[32] Wild, P., and W. R. Gilks, "Adaptive Rejection Sampling from Log-concave Density Functions," *Applied Statistics*, Vol. 42, No. 4, 1993, pp. 701–709.

## APPENDIX A

The file in this appendix may be used in conjunction with the BUGS software in order to conduct Bayesian inference using Gibbs sampling for the model in Example 3.

```
# This is the BUGS file "lpareto.bug".

model lpareto;
const
  cases=25, lambda=5000;
var
  x[cases], y[cases], theta;
data in "lpareto.dat";
inits in "lpareto.in";

{
  for (i in 1:12) {
    y[i]~dpar(theta, lambda) I(5000, 6000);
    x[i]<-y[i]-lambda;
    }
  for (i in 13:20) {
    y[i]~dpar(theta, lambda) I(6000, 7000);
    x[i]<-y[i]-lambda;
    }
  for (i in 21:23) {
    y[i]~dpar(theta, lambda) I(7000, 8000);
    x[i]<-y[i]-lambda;
    }
  for (i in 24:25) {
    y[i]~dpar(theta, lambda) I(8000,);
    x[i]<-y[i]-lambda;
    }
  theta~dgamma(0.001, 0.001);
}
```

## APPENDIX B

The file in this appendix may be used in conjunction with the BUGS software in order to conduct Bayesian inference using Gibbs sampling for the first model in Example 4.

```
# This is the BUGS file “Normal.bug”.

model Normal;
const
   cases=767, classes=133, years=6;
var
   loss[cases], payroll[cases], class[cases],
   x[cases], prec1[cases], theta[classes],
   mu, tau1, tau2,
   y11, y112, y70, y20, y89;
data in “Normal.dat”;
inits in “Normal.in”;

{
   for (i in 1:cases) {
      x[i]<-loss[i] / payroll[i];
      x[i]~dnorm(theta[class[i]], prec1[i]);
      prec1[i]<-tau1*payroll[i];
      }
   tau1~dgamma(0.001, 0.001);

   for (j in 1:classes) {
      theta[j]~dnorm(mu, tau2);
      }
   mu~dnorm(0, 0.001);
   tau2~dgamma(0.001, 0.001);
   y11<-229.83*theta[11];
   y112<-18809.67*theta[112];
   y70<-54.81*theta[70];
   y20<-1315.37*theta[20];
   y89<-79.63*theta[89];
}
```

## APPENDIX C

The file in this appendix may be used in conjunction with the BUGS software in order to conduct Bayesian inference using Gibbs sampling for the second model in Example 4.

```
# This is the BUGS file “Poisson.bug”.

model Poisson;
const
  cases=767, classes=133, years=6;
var
  loss[cases], payroll[cases], class[cases],
  lambda[cases], theta[cases], alpha[classes],
  mu, tau,
  y11, y112, y70, y20, y89;
data in “Poisson.dat”;
inits in “Poisson.in”;

{
  for (i in 1:cases) {
    loss[i]~dpois(lambda[i]);
    lambda[i]<-theta[class[i]]*payroll[i];
    }

  for (j in 1:classes) {
    log(theta[j])<-alpha[j];
    alpha[j]~dnorm(mu, tau);
    }
  mu~dnorm(0, 0.001);
  tau~dgamma(0.001, 0.001);
  y11<-229.83*theta[11];
  y112<-18809.67*theta[112];
  y70<-54.81*theta[70];
  y20<-1315.37*theta[20];
  y89<-79.63*theta[89];
}
```