Simulating a predictive multivariate total claim distribution with Excel VBA*

Rohana S. Ambagaspitiya
Department of Mathematics and Statistics
University of Calgary
2500 University Drive NW
Calgary, AB, T2N1N4
ambagaspucalgary.ca

Abstract

In this paper we propose a model that can be used to analyse correlated claims data. In the process we introduce a class of multivariate generalized Poisson distributions; then we present posterior distributions of its parameters. These distributions are difficult to manipulate, so we employ markov chain monte carlo methods to draw random numbers from posteriors and predictive distributions. We present a collection of EXCEL VBA functions and subrouines to perform the simulations.

Keywords and Phrases

Gamma distribution, Markov Chain Monte Carlo Methods, Multivariate generalized Poisson distribution.

1 Introduction

Let us assume an actuary has collected claims data over a number of periods for number of different classes in one product; the following three tables are an extraction of claims data for three classes over three periods.

*2000 Mathematics Subject Classification: Primary 65C05. Secondary 65C20, 65C60.
Based on the above claims experience we can construct the following table for claim counts across classes over the three periods.

<table>
<thead>
<tr>
<th>Period</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>

It is obvious that for each additional class of claim counts experience we will add one more claim counts column and for each additional period we will add one more row in the claim counts data. From the above data it is clear that we could use the following to model past experience.

1. Let $N_i$ be the number of claims in one period in class $i$ for $i = 1, 2, 3, \ldots, p$.

2. Let $X_{ij}$ be the $j$th claim in the $i$th class for $j = 1, 2, \ldots, N_i$ and $i = 1, 2, 3, \ldots, p$.

With this representation $(N_1, N_2, N_3, \ldots, N_p)$ forms a claim count vector.

We assume that after a preliminary analysis the actuary has drawn the following conclusions:

1. Compound frequency severity model is appropriate for each class.

2. Gamma distribution is a reasonable model for claim sizes.

3. Poisson distribution is not an appropriate model for modeling claim counts for some classes.

4. There is a positive correlation among claim counts across classes.

These assumptions are very strong and the actuary would have completed a large part of the modeling at this point. We are not interested in presenting details of such an analysis as it is peripheral to our discussion. Readers interested in intricate details of such analyses should consult Klugman, Panjer and Willmot (1998).
With this setting we have to choose a multivariate discrete distribution to model claim counts data. There are large number of multivariate discrete distributions to choose from. Johnson et al. (1997) is a book length account of such distributions. Also, one can use Copulas to create multivariate distributions with given marginals as prescribed in Nelson (2006). However, in general multivariate discrete distributions are complicated and difficult to manipulate as opposed to multivariate continuous distributions. Also, we know very little about parameter estimation of multivariate discrete distributions. It has been shown that the generalized Poisson distribution is a good alternative to Poisson distribution in many situations; see for example Consul (1987). Therefore, in this paper we use a form of multivariate generalized Poisson distribution to model claim counts. We must emphasize that the actuary needs to compare a few different models before choosing one; the purpose of this paper is not to describe model selection, it is to facilitate the use the multivariate generalized Poisson distribution for modeling claim counts.

In this paper we present a method for simulating individual claims conditioned on the experience; i.e. we will present a technique to simulate predictive distribution. We do it as a two stage process. We first simulate multivariate observations from the predictive distribution of the claim counts for the next period. Then we simulate the posterior parameters of the individual claim sizes distribution. Then we simulate multivariate total claims. We use Markov Chain Monte Carlo (MCMC) method to simulate predictive distributions. We implement algorithms in Visual Basic for Applications (VBA 2003), so that users can integrate them in their applications as macros in Excel.

In Section 2 we present details necessary to carry out the MCMC simulation for posterior distribution of claim sizes. In Section 3 we first introduce the multivariate generalized Poisson distribution. Then we develop posterior distributions of parameters. In Section 4 we present the Excel VBA implementation details and results of our simulation.

2 Gamma distribution for modelling Claim Sizes

We use the gamma distribution in the following form:

\[ f(x | \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x}, \quad x > 0, \quad \alpha > 0, \beta > 0. \]  

Instead of writing the above functional form we may write \( X \sim \text{Gamma}(\alpha, \beta) \) to denote that the random variable \( X \) is distributed as gamma with parameters \( \alpha \) and \( \beta \).
For each of the portfolios we can fit a different distribution, or by aggregating all the claims we could fit one distribution. In either case, without using too many subscript, let us assume the observed claim sizes are \(x_1, x_2, \ldots, x_n\). Then the likelihood function \(L(x|\alpha, \beta)\) takes the form

\[
L(x|\alpha, \beta) = \prod_{i=1}^{n} f(x_i|\alpha, \beta). \tag{2}
\]

Let us assume the prior distribution for each parameter is also gamma, i.e.

\[
\pi_1(\alpha) = \frac{\gamma_1^\delta_1 \alpha^{\delta_1-1}}{\Gamma(\delta_1)} e^{-\gamma_1 \alpha}, \tag{3}
\]

\[
\pi_2(\beta) = \frac{\gamma_2^\delta_2 \beta^{\delta_2-1}}{\Gamma(\delta_2)} e^{-\gamma_2 \beta}, \tag{4}
\]

with \(\gamma_i, \delta_i, i = 1, 2\) known values. Using these priors we could get the following conditional posterior distributions for the parameters \(\alpha\) and \(\beta\)

\[
p(\alpha|x, \beta) \propto L(x|\alpha, \beta)\pi_1(\alpha) \tag{5}
\]

\[
\beta|x, \alpha \sim \text{Gamma}(n\alpha + \delta_2, \sum_{i=1}^{n} x_i + \gamma_2). \tag{6}
\]

Now we can use the MCMC method to draw random numbers from the posterior distribution of \(p(\alpha, \beta|x)\).

### 3 Multivariate generalized Poisson distribution for modelling claim counts

Let us assume \(M_i\), the \(i\)th element of a random column vector \(M\) of size \(q\), is distributed as generalized Poisson with parameters \(\lambda_i, \theta\) for \(i = 1, 2, \ldots, q\); i.e.

\[
\Pr(M_i = m) = f(m|\lambda_i, \theta) = \frac{\lambda_i(\lambda_i + m\theta)^{m-1}}{m!} \exp(-\lambda_i - m\theta), \quad m = 0, 1, 2, \ldots;
\]

where the parameters \(\lambda_i > 0\) and \(0 \leq \theta < 1\). The means and variances of \(M_i\)'s are given by,

\[
E(M_i) = \frac{\lambda_i}{1 - \theta} \tag{7}
\]

\[
Var(M_i) = \frac{\lambda_i}{(1 - \theta)^3}. \tag{8}
\]
Let us assume that $M_1, M_2, \ldots, M_n$ are independent. Note that having a common parameter $\theta$ across all $q$ distributions ensures that the distribution of any binary combination of $M_1, M_2, \ldots, M_p$ is generalized Poisson. This means that if $\mathbf{A}$ is a $p \times q$ binary matrix and if $\mathbf{N}$ is a column vector defined as

$$\mathbf{N} = \mathbf{A}\mathbf{M},$$

is then distributed as multivariate generalized Poisson.

With a little manipulation we can show that the mean vector and the covariance matrix of $\mathbf{N}$ in (9) becomes,

$$E[\mathbf{N}] = \mathbf{A}E[\mathbf{M}],$$
$$COV(\mathbf{N}) = \mathbf{A}\text{D}\mathbf{A}^T,$$

where $\mathbf{D}$ is a diagonal matrix of size $q \times q$ with

$$\text{diag}(\mathbf{D}) = [\text{Var}[M_1], \text{Var}[M_2], \ldots, \text{Var}[M_q]].$$

From these results we can write the mean vector of our multivariate GPD as

$$E[\mathbf{N}] = \mathbf{A}\begin{bmatrix}
\frac{\lambda_1}{(1-\theta)} \\
\frac{\lambda_2}{(1-\theta)} \\
\vdots \\
\frac{\lambda_q}{(1-\theta)} 
\end{bmatrix}.$$  

(12)

The covariance matrix can be obtained by substituting

$$\text{diag}(\mathbf{D}) = \{\frac{\lambda_1}{(1-\theta)^3}, \frac{\lambda_2}{(1-\theta)^3}, \ldots, \frac{\lambda_q}{(1-\theta)^3}\},$$

in (11). Since all the elements in $COV(\mathbf{N})$ are non-negative, the correlation between elements of $\mathbf{N}$ are positive. To use this multivariate generalized distribution we need to specify the matrix $\mathbf{A}$. We could write

$$\mathbf{A} = [\mathbf{I}_{p \times p}\mid \mathbf{B}_{p \times (q-p)}],$$

(14)

with $\mathbf{I}$ as the identity matrix of dimension $p \times p$; $q$ could take any value in the range $p+1$ to $2^p - 1$. The matrix $\mathbf{B}$ is a binary matrix of dimension $p \times (q-p)$. If $q = 2^p - 1$ the matrix takes special form. The first column is $[1, 1, \ldots, 1]^T$; the next $pC_2$ columns contain exactly two elements of 1 and 0 in other places; these followed $pC_3$ columns containing exactly 3 elements of 1 and 0 in other places. In general $pC_j$ columns followed by $\sum_{i=2}^{j-1} pC_i$ columns of $\mathbf{B}$ contains
\[ j \text{ elements of 1 and 0 in other places for } j = 2 \text{ to } j = p - 1. \] For values of \( q < 2^p - 1 \) we may take a subset of columns of \( B_{p \times (2^p - 1 - p)} \).

Note that \( B \) as a \( p \times (2^p - 1 - p) \) matrix was first suggested in Teicher (1954) for defining the multivariate Poisson distribution. Since then many have used the form with only the first column of \( B \), for example Prekopa and Szantai (1978) used it to generate multivariate gamma distribution; Vernic (1997) used it to generate bivariate generalized Poisson distribution.

Let us write \( M_1 \) for the column vector containing the first \( p \) elements of \( M \), and \( M_2 \) for the column vector containing the remaining elements. Thus,

\[
N = [I_{p \times p} | B_{p \times (q - p)}]\begin{bmatrix} M_{1p \times 1} \\ M_{2(q - p) \times 1} \end{bmatrix}. \tag{15}
\]

We can simplify (15) as

\[
N = M_1 + BM_2 \\
M_1 = N - BM_2.
\]

From this we can write the multivariate probability function of \( N \) as

\[
\Pr[N = n | \Theta] = \sum_{m'} \Pr[M_1 = n - Bm'] \Pr[M_2 = m']
= \sum_{m'} \prod_{i=1}^{p} f(n_i - B_i m'_i | \lambda_i, \theta) \prod_{i=p+1}^{q} f(m_{i-p} | \lambda_i, \theta), \tag{16}
\]

where the summation is over all possible values of \( m' = [m'_1, m'_2, \ldots, m'_{q-p}]^T \) with the restriction \( n_i - B_i m'_i \geq 0 \) for all \( i = 1, 2, \ldots, p \) and \( B_i \) is the \( i \)th row of \( B \). Also the parameter vector \( \Theta = (\lambda_1, \lambda_2, \ldots, \lambda_q, \theta) \) contains \( q + 1 \) parameters. Let us assume the observed multivariate claim count sample over \( k \) periods is the \( p \times k \) matrix \( n \) with \( n = [n_1, n_2, \ldots, n_k] \). Then the likelihood function can be written as

\[
L(n | \Theta) = \prod_{t=1}^{k} \Pr[N = n_t | \Theta]. \tag{17}
\]

Once we substitute the multivariate probability function given in (16), we see that the likelihood function takes a very complicated form. Let us write \( \pi_1(\lambda_i) \) for the prior pdf of \( \lambda_i \) for \( i = 1, 2, \ldots, q \) (i.e. they all have the same functional form) and \( \pi_2(\theta) \) for the prior of \( \theta \). Then the posterior distribution of the parameter \( p(\Theta | n) \) will take the form,

\[
p(\Theta | n) \propto \prod_{t=1}^{k} \Pr[N = n_t | \Theta] \prod_{i=1}^{q} \pi_1(\lambda_i) \pi_2(\theta). \tag{18}
\]
It can easily be seen that manipulating this posterior distribution is difficult due to the form of the likelihood function. However, if we consider $M_2$ as parameters then we could look at the posterior distribution of the parameters which are conditionally independent. This technique is called data augmentation in Bayesian literature. Let us assume $[m'_1, m'_2, \ldots, m'_k]$ is the augmented sample (parameter vector) where each element is a column vector of size $q - p$. In this case the posterior distributions of parameters would take a slightly different form.

\[
p(\lambda_i | n, m', \theta) \propto \prod_{t=1}^{k} f(n_{it} - B_i m'_t | \lambda_i, \theta) \pi_1(\lambda_i), \quad i = 1, 2, \ldots, p \tag{19}
\]

\[
p(\lambda_i | n, m', \theta) \propto \prod_{t=1}^{k} f(n'_{it} | \lambda_i, \theta) \pi_1(\lambda_i), \quad i = p + 1, p + 2, \ldots, q \tag{20}
\]

\[
p(\theta | n, m', \Theta / \theta) \propto \left\{ \prod_{i=1}^{p} \prod_{t=1}^{k} f(n_{it} - B_i m'_t | \lambda_i, \theta) \right\} \left\{ \prod_{i=p+1}^{q} \prod_{t=1}^{k} f(m'_t | \theta_i, \theta) \right\} \pi_2(\theta) \tag{21}
\]

\[
p(m'_t | n, \Theta) \propto \prod_{i=1}^{p} f(n_{it} - B_i m'_t | \lambda_i, \theta) \prod_{j=p+1}^{q} f(m'_{(j-p),t} | \theta_j, \theta), \quad \text{for } t = 1, 2, \ldots, n. \tag{22}
\]

Here $\pi_1(\cdot)$ is the prior distribution of $\lambda_i$ and the $\pi_2(\cdot)$ is the prior of $\theta$.

4 VBA Implementation details

We have implemented four Forms in VBA so that user could navigate through a series of dialogue boxes indicating the desired inputs. The first dialogue box is for the user to specify the data range. The second dialogue box is for choosing the $B$ matrix. The third box allows the user to select parameters for prior distributions. The final box allows the user to indicate the required number of simulations. All the user inputs are checked for validity.

4.1 Implementation of Claim size analysis

With (5) and (6) we implement MCMC method to simulate the joint posterior distribution for $\alpha, \beta$. In the development stage, we realized that when the mean of the gamma distribution is very large the Excel worksheet function GAMMAINV() does not compute the appropriate values. Therefore we use the Metropolis-Hasting (MH) algorithm to simulate from the conditional posterior
distribution \( \beta \) conditioned on \( x \) and \( \alpha \). For the proposal distribution we use the normal distribution with the mean as the value in the preceding iteration and the standard deviation of 0.05 for \( \alpha \) and 0.005 for \( \theta_2 \). We use moment estimates of \( \alpha \) and \( \beta \) as the initial values in MH algorithm; i.e. we use

\[
\alpha = \frac{\bar{x}}{s^2} \\
\beta = \frac{\bar{x}^2}{s^2},
\]

as the initial values; \( \bar{x} \) is the sample mean and \( s \) is the sample standard deviation. In the Excel macros users may change the appropriate prior distribution parameter values \( \gamma_i, \delta_i, i = 1, 2 \). Also users may use numerous charting facilities available in Excel to make sure that simulated parameter sequence is random.

### 4.2 Implementation of claim count analysis

We used (19) to (22) to generate parameters in each step of the MH algorithm. First we wrote Excel function to compute the probability function of the GPD for given parameters. To avoid overflowing the calculation with factorials we computed probabilities recursively. i.e. we used the fact

\[
\frac{\Pr[N = n]}{\Pr[N = n - 1]} = \left( \frac{\lambda_i + n\theta}{\lambda_i + (n - 1)\theta} \right)^{n-2} (\lambda_i + n\theta) \exp(-\theta), \quad \text{for } n = 1, 2, \ldots
\]

with

\[
\Pr[N = 0] = \exp(-\lambda_i).
\]

We choose prior distributions of \( \lambda_i \) to be gamma for \( i = 1, 2, \ldots, q \) and \( \theta \) to be Beta as suggested by Scollnik (1998). The proposal distribution for each parameter is normal with mean as the value of the parameter in the preceding iteration; standard deviation for \( \lambda_i \) is 0.05 for \( i = 1, 2, \ldots, q \) and for \( \theta \) it is 0.005. We wrote numerous procedures and functions to perform calculations in intermediate steps.

In testing, we realize that the implementation of (22) does not work properly. Therefore we just simulated each element \( m'_t \) from the GPD with appropriate parameters and with the added condition that \( n_{it} - B_i m'_t \geq 0 \) for all \( i \).

The initial values for the parameters are obtained through the “Solver” in Excel. These values can be considered as some form of modified moment
estimates. Let us write $\bar{x}$ for the sample mean vector and $S$ for the sample covariance matrix. Then let us define the following two quantities

$$D_1 = \sum_{i=1}^{p} (\bar{x}_i - Mean_i)^2$$

$$D_2 = \sum_{i=1}^{p} \sum_{j=1}^{p} (S_{i,j} - Var(i,j))^2,$$

where $Mean_i$ is the $i^{th}$ element in (12) and $Var(i,j)$ is the $i,j$th element in the covariance matrix for multivariate GPD (i.e. (11)). We call the solver within our VBA modules to solve for the parameters by setting $D_1 = 0$ while minimizing $D_2$.

### 4.3 VBA Implementation of predictive multivariate total claims distribution

This module is the simplest among all three modules. It reads appropriate claim counts to be generated and the posterior parameters of the claim sizes and then it simulates that many claims and computes the total claims. Since all the inputs to this module are generated from previous modules, no error checking is required.

As a test we simulated the total claims distribution with 1000 burn-in simulations followed by 5000 simulations for the hypothetical data set with minimum number of parameters and default prior parameters. The percentiles are given in the following table.

<table>
<thead>
<tr>
<th>Portfolio</th>
<th>Percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>95%</td>
</tr>
<tr>
<td>1</td>
<td>88.03</td>
</tr>
<tr>
<td>2</td>
<td>112.91</td>
</tr>
<tr>
<td>3</td>
<td>134.38</td>
</tr>
</tbody>
</table>

**Acknowledgment**

This research project was supported by a grant from the CKER of the Society of Actuaries.
A Markov Chain Monte Carlo Method Algorithm

In this appendix we briefly describe the implementation of MCMC algorithm in a general setting. For a thorough discussion of MCMC methods and their applications one may refer to standard text books such Gelman et al. (1992) or Gilks et al. (1996). Let us define

\[ X = (x_1, x_2, \ldots, x_n)^T \]

as an observed sample of size \( n \) from the univariate distribution \( f(x|\Theta) \). Here the parameter \( \Theta \) is a vector containing \( p \) elements,

\[ \Theta = (\theta_1, \theta_2, \ldots, \theta_p)^T. \]

We assume the parameters are conditionally independent and the prior distribution of \( \theta_j \) is \( p_j(\theta_j) \), for \( j = 1, 2, \ldots, p \). Therefore the likelihood function \( L(X|\Theta) \) takes the form

\[ L(X|\Theta) = \prod_{i=1}^{n} f(x_i|\Theta). \]

The joint posterior distributions of the parameter vector up to the normalizing constant is

\[ p(\Theta|X) \propto \prod_{i=1}^{n} f(x_i|\Theta) \prod_{j=1}^{p} p_j(\theta_j). \quad (25) \]

Except in a few cases, the posterior distribution in (25) can not be evaluated. However, the MH method prescribes an attractive way to simulate the posterior distribution. We describe the algorithm in the following manner.

Step 1 Specify the initial guess of the parameter vector \( \Theta_0 \) (for example this could be the modes of each of the prior distributions).

Step 2 Specify proposal distributions, \( q_j(\theta_j|\theta) \) for each of the parameters \( \theta_j \), for \( j = 1, 2, \ldots, p \).

Step 3 MCMC Simulation

\[
\text{For } t \text{ from } 1 \text{ to } B + N \text{ do} \\
\text{Let } \Theta^t = [\theta_1^t, \theta_2^t, \ldots, \theta_p^t] \\
\text{For } j \text{ from } 1 \text{ to } p \text{ do}
\]
simulate $\theta^*_j$ from $q_j(\theta|\theta^{t-1})$

Let $\Theta_{t|j} = \{\theta_1^t, \theta_2^t, \ldots, \theta_j^t-1, \theta_j^*, \theta_j^{t-1}, \theta_{j+1}^t, \theta_{j+2}^t, \ldots, \theta_p^t\}$

Let $r = \frac{p(\Theta_{[i]}^t | X)_{\tau_j}(\theta_j^{t-1} | \theta_j^*)}{p(\Theta_{t-1}^t | X)_{\tau_j}(\theta_j^* | \theta_j^{t-1})}$

simulate $U$ from uniform$[0,1]$

set $\theta_j^t = \begin{cases} 
\theta^*_j & \text{if } U < \min(r, 1) \\
\theta_j^{t-1} & \text{otherwise}
\end{cases}$

end do

if $t > B$ then
   simulate $x_{t-B}^{\text{pred}}$ from $f(x | \Theta^t)$
endif
end do

Step 3 Simulate observations from the predictive distributions after a certain number of burn-ins. At the end we could construct the empirical distribution based on the simulated sample $x_1^{\text{pred}}, x_2^{\text{pred}}, \ldots, x_N^{\text{pred}}$ and this would be the predictive distribution.

B References


Software Disclaimer:

Important: The totalclaims.xls ("Software") posted on this site is the property of the Society of Actuaries ("SOA") and is protected under U.S. and international copyright laws. It was created for the SOA by Rohana S. Ambagaspitiya.

The Software has been developed for the benefit of actuaries FOR EDUCATIONAL USE ONLY, although others may find it useful. SOA makes the Software available to individual users for their personal use on a non-exclusive basis. No commercial use, reproduction or distribution is permitted whatsoever.

SOA and the authors make no warranty, guarantee, or representation, either expressed or implied, regarding the Software, including its quality, accuracy, reliability, or suitability, and HEREBY DISCLAIM ANY WARRANTY REGARDING THE SOFTWARE’S MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. SOA and the authors make no warranty that the Software is free from errors, defects, worms, viruses or other elements or codes that manifest contaminating or destructive properties. In no event shall SOA or the authors be liable for any damages (including any lost profits, lost savings, or direct, indirect, incidental, consequential or other damages) in connection with or resulting from the use, misuse, reliance on, or performance of any aspect of the Software including any instructions or documentation accompanying the Software. SOA and the authors make no representation or warranty of non-infringement of proprietary rights of others with respect to the Software. The entire risk as to the uses, outputs, analyses, results and performance of the Software is assumed by the user. This Disclaimer applies regardless of whether the Software is used alone or with other software.