AUSTRIAN JOURNAL OF STATISTICS
Volume 31 (2002), Number 2&3, 221-229

Hierarchical Bayesian Models for Multiple Count Data

Radu Tunaru
Economics Department, London Metropolitan University

Abstract: The aim of this paper is to develop a model for analyzing multiple response models for count data and that may take into account complex correlation structures. The model is specified hierarchically in several layers and can be used for sparse data as it is shown in the second part of the paper. It is a discrete multivariate response approach regarding the left side of models equations. Markov Chain Monte Carlo techniques are needed for extracting inferential results. The possible correlation between different counts is more general than the one used in repeated measurements or longitudinal studies framework.

Keywords: Multiple Count Data, Multivariate Poisson, Hierarchical Bayesian Model, MCMC, Ranking.

1 Introduction

The possible correlation between different counts is not due to time and, in practice, it is a more complex multivariate structure than repeated measurements or longitudinal studies framework. Techniques for modelling multiple counts jointly have not been extensively developed in the statistical literature, mainly because of the lack of multivariate discrete distribution that could support complex correlation structures. Prior Bayesian and EB research related to multiple response variables has been focused on longitudinal studies for clinical trials or biostatistical data (Breslow and Clayton, 1993; Zeger and Karim, 1991; Gilks et al., 1996; Carlin and Louis, 1996).

In this paper a multiple response model for counts is developed. The model is applied to a real-world set of data involving accident frequencies. It is specified hierarchically and belongs to the fully Bayesian family. At the same time the hierarchical model investigated here can be adapted to models other than multivariate count data.

The fitting is made possible by applying adequate Markov Chain Monte Carlo (MCMC) methods. For the applied analysis the WinBUGS package was used, an excellent platform for Bayesian modelling.

2 Bayesian Multivariate Poisson-log Normal Model

In this section, a discrete multivariate distribution is described as a feasible solution for discrete data modelling with multiple responses. The idea is simple, Aitchison and Ho (1989), but powerful MCMC computational methods are needed to put it into practice. For all $k \in \{1, 2, \ldots, N\}$ and $i \in \{1, 2, \ldots, M\}$ we write

$$Y_{ki} \mid \lambda_{ki} \overset{ind}{\sim} \mathrm{Pois}(\lambda_{ki}) \tag{1}$$

$$(\log(\lambda_{k1}), \ldots, \log(\lambda_{kM}))' \mid \mu, T \overset{iid}{\sim} \mathrm{N}_M(\mu, T) \tag{2}$$

where $T = \Sigma^{-1}$ is the precision matrix. The probability density function of the M-dimensional log normal distribution is

$$p(\lambda \mid \mu, T) = (2\pi)^{-\frac{M}{2}} (\prod_{i=1}^{M} \lambda_i)^{-1} |T|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\log\lambda - \mu)' T (\log\lambda - \mu)\right). \quad (3)$$

The multivariate Poisson-log normal distribution, denoted subsequently by $P\Lambda^M(\mu, T)$, is the combination of independent Poisson distributions with multivariate log normal distribution for the Poisson means. The probability density function of $P\Lambda^M(\mu, T)$ is exactly the marginal density of Y's conditioned on μ and T only

$$p(y_1, \ldots, y_M \mid \mu, T) = \int_{\mathbf{R}_+^M} \prod_{i=1}^{M} \mathrm{Pois}(y_i \mid \lambda_i) p(\lambda_i \mid \mu, T) d\lambda_1 \cdots d\lambda_M \quad (4)$$

where $y_1, \ldots, y_M = 0, 1, \ldots$. The important moments of this distribution can be easily calculated. If $\Sigma = (\sigma_{ij})$ then

$$\begin{aligned}
\mathrm{E}(Y_i) &= \mathrm{E}(\mathrm{E}(Y_i \mid \lambda_i)) = \exp(\mu_i + \frac{1}{2}\sigma_{ii}) = a_i && (5)\\
\mathrm{var}(Y_i) &= \mathrm{E}(\mathrm{var}(Y_i \mid \lambda_i)) + \mathrm{var}(\mathrm{E}(Y_i \mid \lambda_i)) && (6)\\
&= a_i + a_i^2(\exp(\sigma_{ii}) - 1) > \mathrm{E}(Y_i) &&\\
\mathrm{cov}(Y_i, Y_j) &= \mathrm{E}(\mathrm{cov}(Y_i, Y_j \mid \lambda)) + \mathrm{cov}(\mathrm{E}(Y_i \mid \lambda_i), \mathrm{E}(Y_j \mid \lambda_j)) && (7)\\
&= \mathrm{cov}(\lambda_i, \lambda_j) = a_i a_j (\exp(\sigma_{ij}) - 1). &&
\end{aligned}$$

Some important immediate consequences are that there is overdispersion for the marginal distributions, and

$$|\mathrm{corr}(Y_{ki}, Y_{kj})| < |\mathrm{corr}(\lambda_{ki}, \lambda_{kj})|, \quad \mathrm{sgn}(\mathrm{corr}(Y_{ki}, Y_{kj})) = \mathrm{sgn}(\mathrm{corr}(\lambda_{ki}, \lambda_{kj}))$$

which are special cases of the results of the mean-variance model. Negative and positive correlations are supported by this mixed distributions, which gives it an advantage over other multivariate discrete distributions such as multinomial or negative multinomial. However, the estimation of the parameters is not straightforward. For maximum likelihood estimation, a re-parameterization and a mixture of Newton-Raphson and steepest ascent methods are helpful but computationally intensive (see Aitchison and Ho, 1989).

In this paper we shall use MCMC methods (Metropolis-Hastings algorithm) to obtain inference summaries about the parameters μ and T. In a fully Bayesian context, further prior distributions, probably non-informative, are required for μ and T. The recommended parametric distributions are normal for μ and Wishart for T (Carlin and Louis, 1996; Gelman et al., 1995).

3 Data Analysis of Multiple Count Data

In this section the model introduced in the previous section is applied to a multiple count dataset. The counts represent road accidents. The ability to model joint responses provides another dimension to statistical modelling. The advantage of using MCMC techniques is that the same model output can be used to provide inference on several problems

Table 1: Total number of accidents for each category of accidents

Severity	Number of vehicles involved	Total number of accidents
fatal or serious	1	443
	2 or more	852
slight	1	796
	2 or more	2160

like model selection, goodness-of-fit, ranking the units of the analysis according to different criteria, and so on.

3.1 Data

The data includes all accidents between 1984 and 1991 and the units of the analysis are 156 single carriageway link sites in Kent. The links are defined as road sections between two major junctions, or between changes in carriageway type (single or dual), or between changes in speed limits.

Because manual counts can be sparse in both location and time a simple linear regression was used to fill in the missing years and account for some of the variation in individual counts.

This set of data was provided by the Transport Management Research Centre at Middlesex University. The number of accidents at each site was disaggregated by accident severity, having two levels KSI = fatal or serious and S = slight, and the number of vehicles involved, with two levels, 1 vehicle and 2 or more vehicles. The cross-classification gives four possibly correlated groups of observations. The observed number of accidents in each group is provided in Table 1 where it can be seen that there are sites with zero accidents for the total number of accidents and sites with zero counts for some types of accident.

3.2 Statistical Inference Based on MCMC

The inference process is based on sample of 10000 values either after a burn-in of 40000 iterations from a single chain or after a burn in of 20000 iterations from two chains started from dispersed initial points. The Brooks-Gelman-Rubin statistic, as calculated in WinBUGS was less than 1.05 for all parameters of the models investigated. However, even after checking that the Markov chain has converged from an empirical point of view and parameters are reliably estimated it is necessary to check the goodness-of-fit of the model before applying the results. This was done using Bayesian P-values, see Carlin and Louis

(1996), for the Pearson χ^2 measure of discrepancy that is generally used in this area of modelling.

3.3 Bayesian Inference

This section contains the results when applying the model as a hierarchical Bayesian model. The model reveals qualitative and quantitative relationships between the counts of different type.

The combination of a Poisson distribution with a multivariate log normal distribution was described above by equation (4), as a discrete multivariate distribution for modelling multiple counts. Starting from this multivariate Poisson-log normal distribution a hierarchical, fully Bayesian model is proposed as

$$\begin{aligned} Y_{ki} \mid \lambda_{ki} &\overset{ind}{\sim} \mathrm{Pois}(\lambda_{ki} \\ (\log(\lambda_{ki}))_{i=1,\dots,4} \mid \mu, T &\overset{iid}{\sim} \mathrm{N}_4(\mu, T) \\ \mu_i &\overset{iid}{\sim} \mathrm{N}(0, 0.0001) \\ T &\sim \mathrm{Wishart}(R, 4) \end{aligned} \tag{8}$$

where $\mathrm{N}_M(\mu, T)$ is the M-dimensional multivariate normal distribution with mean vector μ and with T the inverse of the covariance matrix. The hyper-prior parameters R and $\pi \geq M$ are known, usually taking $\pi = M$ for vague priors. The parameterisation of the Wishart probability density function is

$$f(\mathbf{X} \mid R, \pi) \propto |R|^{\frac{\pi}{2}} |\mathbf{X}|^{\frac{\pi - M - 1}{2}} e^{-\frac{1}{2}\mathrm{Tr}(R\mathbf{X})}$$

following Spiegelhalter et al. (1996). The Wishart prior is used for the inverse of the covariance matrices of multivariate normal distributions and because $\mathrm{E}(\mathbf{X}) = \pi R^{-1},\ \ R^{-1}$ is best interpreted as the expected prior precisions of the random effects μ. Small values of π correspond to vaguer prior distributions and it is recommended (Spiegelhalter et al., 1996) to take $\pi = M$.

The Bayesian P values for the Pearson χ^2 measure of discrepancy, for each type of accident, are 0.87 for KSI accidents with one vehicle, 0.77 for KSI accidents with two or more, 0.71 for S accidents with one vehicle and 0.62 for S accidents with two or more vehicles. It can be concluded that the data does not contradict the model so the inferences are reliable.

The posterior estimates of the parameters of interest for the multivariate Poisson-log normal model is given in Table 2. The covariance matrix $\Sigma = T^{-1}$ is provided because it makes a straightforward link with possible covariance structure of the observed data. The matrix given by (9)

$$T = \begin{pmatrix} 4.42 & -1.55 & -1.65 & -0.55 \\ -1.55 & 3.76 & -0.99 & -1.52 \\ -1.65 & -0.99 & 3.18 & -0.36 \\ -0.55 & -1.52 & -0.36 & 2.44 \end{pmatrix} \tag{9}$$

contains the posterior means of the elements of T, the inverse covariance matrix. There are weak partial correlations between KSI accidents with 1 vehicle and slight accidents

Table 2: Posterior estimation of parameters of multivariate Poisson-log normal model.

parameter	mean	sd	2.5%	97.5%
σ_{11}	2.15	0.41	1.46	3.09
σ_{12}	2.34	0.41	1.65	3.29
σ_{13}	2.10	0.39	1.48	3.00
σ_{14}	2.25	0.38	1.62	3.11
σ_{21}	2.34	0.41	1.66	3.29
σ_{22}	3.04	0.54	2.16	4.27
σ_{23}	2.48	0.45	1.78	3.54
σ_{24}	2.78	0.45	2.04	3.79
σ_{31}	2.10	0.39	1.48	3.00
σ_{32}	2.48	0.45	1.78	3.54
σ_{33}	2.49	0.47	1.75	3.61
σ_{34}	2.39	0.41	1.73	3.32
σ_{41}	2.25	0.38	1.62	3.11
σ_{42}	2.78	0.45	2.04	3.79
σ_{43}	2.39	0.40	1.73	3.32
σ_{44}	3.04	0.47	2.25	4.09
μ_1	0.28	0.15	-0.03	0.56
μ_2	0.67	0.16	0.34	0.97
μ_3	0.79	0.16	0.46	1.08
μ_4	1.65	0.15	1.35	1.94

with 2+ vehicles, between KSI accidents with 2+ vehicles and slight accidents with 1 vehicle and between slight accidents with 1 vehicle and slight accidents with 2+ vehicles.

3.4 Ranking the Sites

Identifying hazardous sites is vital since large amounts of money can be wasted just because the dangerous sites have not been identified as such. This is linked to the problem of ranking the sites according to the severity they posed in terms of number of accidents and severity.

3.4.1 Ranking by the Probability that a Site is the Worst

The posterior probability that the site k is worse than all the others by a factor of v, for accident type i, is

$$p_{ki}(v) = \Pr(\lambda_{ki} > v\lambda_{ji} \text{ for all } j \neq k \mid y)$$

where $v > 0$. For example, when $v = 1$ this is the probability that the site is the worst one. The factor v should be established prior to the analysis by the practitioner. The posterior probability that is used as a criterion for ranking represents a measure of how much worse one accident site is compared with all the others. In practice arbitrarily selected υ-values like $\upsilon = 1, 1.1, 1.25$ are used.

Only the sites with corresponding probabilities larger than 10^{-4} are presented in Table 3 summarising the results.

By the measure studied in this section, it seems that there are not many dangerous sites for slight accidents with only one vehicle. One reason might be that site 90 is so bad that almost the whole probability is concentrated on this site, and there is not very much left to distinguish between the others. This site is particularly interesting. It is the urban link that runs along the sea front at the resort of Margate. Thus, there would be a high volume of holiday makers both pedestrians and drivers. The high pedestrian flow distinguishes it from the other links and special safety measures need to be implemented.

Table 3 corresponds to the first ranking criterion looking at the probability that a unit is the worst. This can be used for long term projects. The practitioner then can see different lists and make an ad-hoc decision accordingly. The point to bear in mind is that the list of selected sites should not contain just a few sites or too many sites. The value $\upsilon = 1$ is always a good start and depending on the results obtained, the practitioner can modify υ accordingly. When $\upsilon = 1$ it is true that

$$\sum_k p_{ki}(1) = 1$$

and this is convenient for checking that the calculations are correct.

3.4.2 Ranking by Posterior Distributions of Ranks

The second criterion for ranking sites investigated here is based on the ranks r_{ki} of the mean parameters λ_{ki} which are the site specific parameters. The ranking process can be done again for each type of accident i, but only the result for the first type of accidents are shown here. The posterior means $\mathrm{E}(\lambda_{ki} \mid y)$ are optimal estimates when the aim is to produce inference about λ_{ki}. However, if the ranks of λ_{ki} are of interest, the conditional expected ranks (or a discretized version of them when they are not integers) are optimal. It is known that ranking the observed data or even the posterior means can perform poorly (Laird and Louis, 1989; Morris and Christiansen, 1996). Consequently, this ranking method is developed using the posterior distribution of the ranks, that is $p(r \mid y)$, and not the posterior distribution $p(\lambda \mid y)$. This differs than the approach proposed by Schluter et al. (1997).

Ranks are notoriously uncertain and it is useful to know the uncertainty associated with them. The approach followed here easily calculates the corresponding posterior probability confidence intervals of the estimated ranks. The ranks will be estimated by the posterior medians, mainly because they are easier to calculate. For each model and each accident type, the posterior median ranks and the associated $2.5\% - 97.5\%$ credible intervals are plotted together for comparison. Sites with ranks to the far right are more dangerous and sites with ranks to the far left are more safe.

The plot in Figure 1 illustrates the estimated statistics of the ranks of lambda's for the first type of accidents. The ranks are ordered and a leaf type shape can be noticed. In addition, sites with the lowest and, respectively highest, rank values, have quite small credible intervals.

Table 3: Ranking probabilities of site with accidents.

Type 1 Site No	Pr	Type 2 Site No	Pr	Type 3 Site No	Pr	Type 4 Site No	Pr
		4	0.0018				
11	0.0020	11	0.1252			11	0.0004
		12	0.1194			12	0.1054
14	0.0076	14	0.2732	14	0.0008	14	0.0626
23	0.0002						
		24	0.0018			24	0.0220
38	0.0002						
41	0.1332	41	0.0014	41	0.0010	41	0.2144
42	0.0004						
46	0.0280	46	0.3228	46	0.0036		
76	0.0004	76	0.0032			76	0.0022
77	0.0058						
90	0.7934	90	0.0064	90	0.9946	90	0.5688
91	0.0004						
95	0.0058						
		98	0.0348			98	0.0206
		102	0.0028				
118	0.0008	118	0.1070				
143	0.0218						

4 Conclusion

The multivariate normal distribution provides a sound base for statistical modelling of multivariate continuous data. In spite of that, for multivariate counts, there is a lack of discrete multivariate distributions that could play the role of Poisson distribution in the univariate case. A consequence is that sometimes inappropriate methods employing continuous multivariate distributions are proposed in order to support a complex structure. The study of Amis (1996) is an example of a good applied statistical work that can be further improved by applying the hierarchical Bayesian methodology proposed here.

In this paper we developed a multiple response model for counts that could support complex correlation structures. This model has been fully Bayesian specified hierarchically in several stages. MCMC algorithms like the Gibbs sampler or the Metropolis-Hastings can be used to obtain inferential results even for sparse data like road accidents data.

The model has been used on a real-world set of data for ranking of units. The ranking of observational units has been done according to two different criteria, one that can be used for long term projects and looking at the probability that a unit is the worst and the other that can be used for short term projects and looking at the ranks of the expected number of counts.

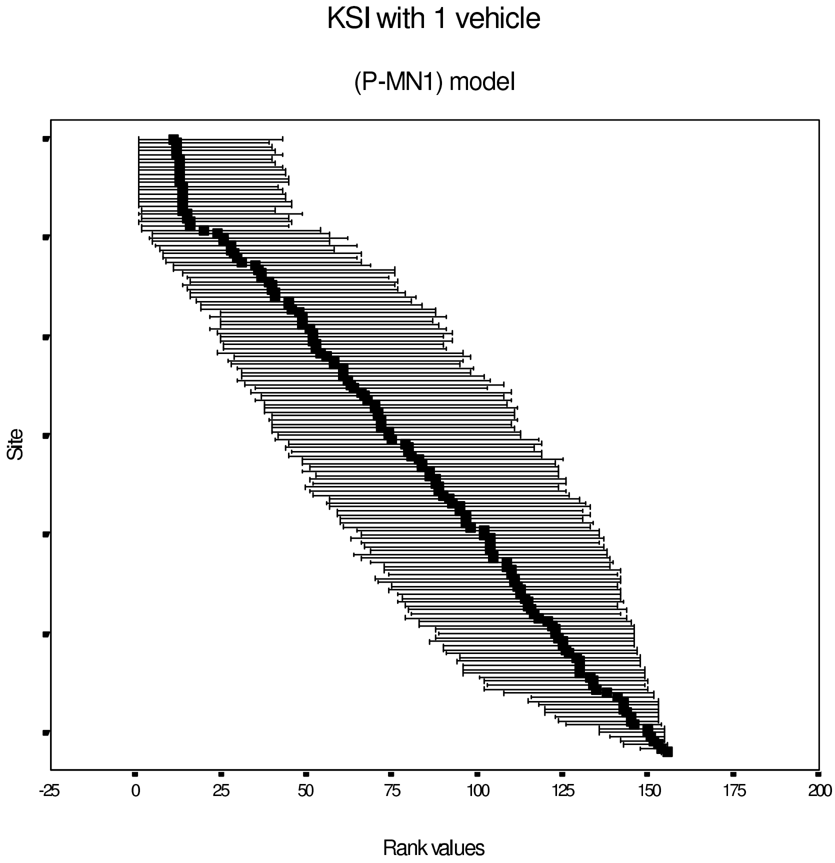

Figure 1: Ordered ranks for first type of accidents and multivariate Poisson-log normal model

The inference process is quite complex and Markov Chain Monte Carlo methods were needed. However, there is a bonus in that once the model has been fitted many interesting questions were answered using the same output.

References

J. Aitchison and C.H. Ho. The multivariate Poisson-log normal distribution. *Biometrika*, 76(4):643–653, 1989.

G. Amis. An application of generalised linear modelling to the analysis of traffic accidents. *Traffic Eng. and Control*, 1996.

N.E. Breslow and D.G. Clayton. Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88:9–25, 1993.

B.P. Carlin and T.A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London, 1996.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.

W.R. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, 1996.

N. Laird and T.A. Louis. Empirical Bayes ranking methods. *Journal of Educational Statistics*, 14:29–46, 1989.

C.N. Morris and C.L. Christiansen. Hierarchical models for ranking and for identifying extremes, with applications. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 5*, pages 277–296, Oxford, 1996. Oxford University Press.

P.J. Schluter, J.J. Deely, and A.J. Nicholson. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. *The Statistician*, 46(3):293–316, 1997.

D.J. Spiegelhalter, A. Thomas, and N.G. Best. Computation on bayesian graphical models. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian statistics 5*, pages 407–425, Oxford, 1996. Oxford University Press.

S.L. Zeger and M.R. Karim. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86, 1991.

Author's address:

Dr. Radu Tunaru
London Metropolitan University
Economics Department
31 Jewry Street
London EC3N 2EY

E-mail: r.tunaru@londonmet.ac.uk