Markov Chain Monte Carlo Analysis of Correlated Count Data

Siddhartha CHIB
John M. Olin School of Business, Washington University, St. Louis, MO 63130 (*chib@olin.wustl.edu*)

Rainer WINKELMANN
IZA Bonn, 53072 Bonn, Germany (*winkelmann@iza.org*)

This article is concerned with the analysis of correlated count data. A class of models is proposed in which the correlation among the counts is represented by correlated latent effects. Special cases of the model are discussed and a tuned and efficient Markov chain Monte Carlo algorithm is developed to estimate the model under both multivariate normal and multivariate-t assumptions on the latent effects. The methods are illustrated with two real data examples of six and sixteen variate correlated counts.

KEY WORDS: Latent effects; Metropolis–Hastings algorithm; Multivariate count data; Poisson–lognormal distribution.

A large literature on the analysis of count data is now available (Cameron and Trivedi 1998, Winkelmann 2000), but only a small portion of it deals with correlated counts. Correlated counts typically arise in three varieties—as genuine "multivariate" data on several related counted outcomes, as longitudinal measurements on a large number of subjects over a short period of time, or as measurements on a small set of subjects over a long period of time (the seemingly unrelated case). Although the longitudinal situation has been actively studied (e.g., see Hausman, Hall, and Griliches 1984; Blundell, Griffith, and Van Reenen 1995; Wooldridge 1997; Chib, Greenberg, and Winkelmann 1998, henceforth CGW) and a number of useful models and approaches are available, the other cases have been analyzed only under simplifying assumptions (King 1989; Jung and Winkelmann 1993; Gurmu and Elder 1998; Munkin and Trivedi 1999). The latter approaches either do not allow a general correlation structure or are difficult to extend beyond the case of a few outcomes.

This article is an effort to deal with both problems. To model the correlation among a large number of counts in a flexible fashion, we introduce a set of correlated latent effects, one for each subject and outcome. Conditioned on the latent effects, the counts are assumed to be independent Poisson with a conditional mean function that depends on the latent effects and a set of covariates. To complete the model we assume that the latent effects follow a multivariate Gaussian distribution with a zero mean vector and full unrestricted covariance matrix. As an extension of this model, we also consider the case in which the latent effects follow a multivariate-t distribution. To estimate this model, we develop a Markov chain Monte Carlo (MCMC) simulation method that is based on the work of CGW. Under this framework, we are able to sample the posterior distribution of the parameters and latent effects without computing the likelihood function of the model.

The methods that we develop in this article can be applied to datasets with large numbers of correlated counts. We demonstrate this feature by fitting our model to a problem with 16 response variables. In our view this is an important illustration that highlights what is possible from a Bayesian simulation-based perspective.

The rest of the article is organized as follows. In Section 1 we present the basic model and some special cases and extensions. The fitting algorithm is developed in Section 2, while Section 3 gives two real data examples. Section 4 concludes.

1. MODEL

Following the usual notation for multivariate data, let $y_i = (y_{i1}, \ldots, y_{iJ})$ denote the collection of J counts on the ith subject in the sample, $i \le n$. Let $b_i = (b_{i1}, \ldots, b_{iJ})$ denote a set of J subject and outcome-specific latent effects, and suppose that, conditioned on b_i and parameters $\beta_j \in R^{k_j}$, the counts y_{ij}, $j \le J$, are independent Poisson:

$$\begin{aligned} y_{ij}|b_i, \beta_j &\sim \text{Poisson}(\mu_{ij}) \\ \mu_{ij} &= \exp(x'_{ij}\beta_j + b_{ij}) \\ &\text{for} \quad j \le J \quad \text{and} \quad i \le n, \end{aligned} \tag{1}$$

where x_{ij} are covariates. To model the correlation among the counts, let

$$b_i|D \sim \text{N}_J(0, D), \qquad i \le n, \tag{2}$$

where D is an unrestricted covariance matrix.

To understand some of the features of this model, let $v_{ij} = \exp(b_{ij})$ and $v_i = (v_{i1}, \ldots, v_{iJ})$. Then $v_i \sim \text{LN}_J(\mu, \Sigma)$, a multivariate lognormal distribution with mean $\mu = \exp(0.5\ \text{diag}(D))$ and dispersion matrix $\Sigma = (\text{diag}(\mu))[\exp(D) - \mathbf{1}\mathbf{1}'](\text{diag}(\mu))$. Hence, $y_{ij}|\lambda_{ij}, v_{ij} \sim \text{Poisson}(\lambda_{ij}v_{ij})$, where $\lambda_{ij} = \exp(x'_{ij}\beta_j)$. This is, therefore, in the form of a Poisson–lognormal distribution as discussed by Aitchison and Ho (1989).

In this setup, the expectation and variance of the marginal joint distribution of y_i can be derived without integration. Let $\tilde{\lambda}_{ij} = \lambda_{ij}\mu$ (i.e., $\tilde{\lambda}_{ij}$ and λ_{ij} differ only by a constant factor), $\tilde{\lambda}_i = (\tilde{\lambda}_{i1}, \ldots, \tilde{\lambda}_{iJ})$, and $\tilde{\Lambda}_i = \text{diag}(\tilde{\lambda}_i)$. Applying the

Journal of Business & Economic Statistics
October 2001, Vol. 19, No. 4

law of the iterated expectation, one obtains $\mathrm{E}(y_i|\beta, D) = \tilde{\lambda}_i$ and $\mathrm{var}(y_i|\beta, D) = \tilde{\Lambda}_i + \tilde{\Lambda}_i[\exp(D) - \mathbf{1}\mathbf{1}']\tilde{\Lambda}_i$, where we have $\beta = (\beta_1, \ldots, \beta_J)$. Hence, the covariance between the counts is represented by the terms

$$\begin{aligned}\mathrm{cov}(y_{ij}, y_{ik}) &= \tilde{\lambda}_{ij}(\exp(d_{jk}) - 1)\tilde{\lambda}_{ik} \\ &= \lambda_{ij}\exp(0.5d_{jj})(\exp(d_{jk}) - 1)\lambda_{ik}\exp(0.5d_{kk}), \\ &\qquad j \neq k,\end{aligned}$$

which can be positive or negative depending on the sign of d_{jk}, the (j, k) element of D. Moreover, the model allows for overdispersion, a variance in excess of the expectation, as long as $d_{jj} > 0$. The correlation structure of the counts is thus unrestricted. Note, however, that the marginal distribution of the counts y_i cannot be obtained by direct computation, requiring as it does the evaluation of a J-variate integral of the Poisson distribution in (1) with respect to the distribution of b_i.

It is interesting to note that our model is similar to that of Gurmu and Elder (1998) except that in their model the distribution of b_{ij} is left unspecified. Under that assumption, the model becomes computationally intractable for anything more than a few correlated counts. As we show in this article, it is possible to fit higher-dimensional models provided one is willing to make a parametric distributional assumption for b_i, which in turn provides a clean interpretation for the correlation structure. The assumption of normality is not crucial and can be generalized. For example, it is easy to let the distribution of the latent effects be multivariate-t instead of the multivariate-normal, as will be discussed, or to model the distribution by a finite mixture of normal distributions. More importantly, it is possible to relax the assumption, implicit in the preceding formulation, that the b_i are independent of the covariates by letting the mean of b_i be a function of one or more of the available covariates. The estimation approach that we will present needs to be modified only slightly to incorporate this feature. Finally, our model can be specialized to the panel-data setting (where the index j represents time) by letting the conditional mean function be $\theta_{ij} = \exp(x_{ij}'\beta + w_{ij}'b_i)$, where w_{ij} is a set of covariates that are a subset of x_{ij}. This is exactly the model of CGW that in turn is a generalization of the model of Hausman et al. (1984). It should be noted that, in this specialization of the general model, fewer than J latent effects appear in the conditional mean function of subject i.

2. ESTIMATION OF THE MODEL

2.1 Likelihood Function

Let us suppose that the observations $y_i = (y_{i1}, \ldots, y_{iJ})$ are conditionally independent across clusters. Then, the likelihood function is the product of the contributions $p(y_i|\beta, D)$, where $p(y_i|\beta, D)$ is the joint probability of the J counts in cluster i given by

$$p(y_i|\beta, D) = \int \prod_{j=1}^{J} f(y_{ij}|\beta_j, b_{ij})\phi_J(b_i|0, D)\mathrm{d}b_i, \qquad (3)$$

where f, as previously, is the Poisson mass function conditioned on (β_j, b_{ij}) and ϕ is the J-variate normal density function. This multiple integral cannot be solved in closed form for arbitrary D, but some simplifications are possible if D is assumed to be a diagonal matrix. To deal with the general case, however, it is necessary to turn to simulation-based methods.

2.2 MCMC Implementation

The main idea of the estimation approach is to focus on the posterior distribution of the parameters and the latent effects and then to summarize this posterior distribution by MCMC methods. Since much has been written about MCMC methods (e.g., see Tierney 1994; Chib and Greenberg 1995, 1996), we can be brief.

With MCMC methods, one designs an ergodic Markov chain with the property that the limiting invariant distribution of the chain is the posterior density of interest. Then, draws furnished by sampling the Markov chain, after an initial transient or burn-in stage, can be taken as approximate correlated draws from the posterior distribution. This output forms the basis for summarizing the posterior distribution and for computing Bayesian point and interval estimates. Ergodic laws of large numbers for Markov chains on continuous state spaces are used to justify that these estimates are simulation consistent, converging to the posterior expectations as the simulation sample size becomes large.

One standard method for constructing a Markov chain with the correct limiting distribution, is via a recursive simulation of the so-called full conditional densities—that is, the density of a set or block of parameters, given the data and the remaining blocks of parameters. Each of the full conditional densities in the simulation is then sampled either directly (if the full conditional density belongs to a known family of distributions) or by utilizing a technique such as the Metropolis–Hastings (M–H) method. An important and crucial point is that these methods do not require knowledge of the intractable normalizing constant of the posterior distribution.

In the present case, we apply MCMC methods to simulate the augmented posterior distribution of the parameters and the latent effect. For the prior on the parameters, assume that (β, D^{-1}) independently follow the distributions $\beta \sim \mathrm{N}_k(\beta_0, B_0^{-1})$, $D^{-1} \sim \mathrm{Wishart}(\nu_0, R_0)$, where $(\beta_0, B_0, \upsilon_0, R_0)$ are known hyperparameters and $\mathrm{Wishart}(\cdot, \cdot)$ is the Wishart distribution with ν_0 df and scale matrix R_0. Then, by Bayes theorem, the posterior density is proportional to

$$\phi_J(\beta|\beta_0, B_0^{-1})f_{\mathrm{W}}(D^{-1}|\nu_0, R_0)\prod_{i=1}^{n} p(y_i|\beta, b_i)\phi_J(b_i|0, D),$$

where f_{W} is the Wishart density. We now consider a sampling procedure to simulate this density.

Following CGW, we construct our Markov chain using the blocks of parameters $\{b_i\}$, β, and D and the full conditional distributions

$$[b|y, \beta, D]; \quad [\beta|y, b]; \quad [D^{-1}|b], \qquad (4)$$

where $b = (b_1, \ldots, b_n)$. The simulation output is obtained by recursively simulating these distributions, using the most recent values of the conditioning variables at each step.

2.3 Sampling b

The target density is $\pi(b|y,\beta,D)=\prod_{i=1}^{n}\pi(b_i|y_i,\beta,D)$, which factors into the product of n independent terms. To sample the ith density,

$$\begin{aligned}\pi(b_i|y_i,\beta,D)&=c_i\phi_J(b_i|0,D)\prod_{j=1}^{J}\exp[-\exp(x_{ij}'\beta_j+b_{ij})]\\&\quad\times[\exp(x_{ij}'\beta_j+b_{ij})]^{y_{ij}}\\&\equiv c_i\pi^+(b_i|y_i,\beta,D),\end{aligned}\tag{5}$$

we utilize the M–H algorithm. As shown by CGW, the proposal density is found by approximating the target density around the modal value by a multivariate-t distribution. Let $\hat{b}_i=\arg\max\ln\pi^+(b_i|y_i,\beta,D)$ and $V_{b_i}=(-H_{b_i})^{-1}$ be the inverse of the Hessian of $\ln\pi^+(b_i|y_i,\beta,D)$ at the mode $\hat{b}_i$. To find these quantities, we use the Newton–Raphson algorithm with the gradient vector $g_{b_i}=-D^{-1}b_i+[y_i-\exp(x_i\beta+b_i)]$ and Hessian matrix $H_{b_i}=-D^{-1}-\operatorname{diag}\{\exp(x_i\beta+b_i)\}$. In practice, three or four steps of the Newton–Raphson algorithm are sufficient to locate the mode of the target density. Then, our proposal density is taken to be $q(b_i|y_i,\beta,D)=f_T(b_i|\hat{b}_i,V_{b_i},\nu)$, a multivariate-$t$ density with ν df where ν is a tuning parameter. We now draw a proposal value b_i^* from $q(b_i|y_i,\beta,D)$ and move to b_i^* from the current point b_i with probability

$$\alpha(b_i,b_i^*|y_i,\beta,D)=\min\left\{\frac{\pi^+(b_i^*|y_i,\beta,D)q(b_i|y_i,\beta,D)}{\pi^+(b_i|y_i,\beta,D)q(b_i^*|y_i,\beta,D)},1\right\}.\tag{6}$$

If the proposal value is rejected, then the next item in the chain is the current value b_i.

2.4 Sampling β

We next sample β given (b,D) from the density that is proportional to

$$\begin{aligned}\pi^+(\beta|y,b,D)&=\phi_k(\beta|\beta_0,B_0^{-1})\prod_{i=1}^{n}\prod_{j=1}^{J}\exp[-\exp(x_{ij}'\beta_j+b_{ij})]\\&\quad\times[\exp(x_{ij}'\beta_j+b_{ij})]^{y_{ij}}\\&=\phi_k(\beta|\beta_0,B_0^{-1})\prod_{j=1}^{J}p(y_{\cdot j}|\beta_j,b_{\cdot j}),\end{aligned}\tag{7}$$

where

$$\begin{aligned}p(y_{\cdot j}|\beta_j,b_{\cdot j})&=\prod_{i=1}^{n}\exp[-\exp(x_{ij}'\beta_j+b_{ij})]\\&\quad\times[\exp(x_{ij}'\beta_j+b_{ij})]^{y_{ij}}\end{aligned}$$

is the mass function of the observations $y_{\cdot j}=(y_{1j},\ldots,y_{nj})$ given β_j and $b_{\cdot j}=(b_{1j},\ldots,b_{nj})$. The factorization given previously utilizes the fact that the counts are conditionally independent given the latent effects. There are two ways to sample this density. One way is to sample β in one block by the M–H algorithm, and the second way is to sample each of the β_j's by a sequence of M–H steps. When the dimension of β is large, as in our first application, the latter approach would be preferred.

To sample β in one block, we develop a tuned proposal density in a manner analogous to that of b_i. The proposal distribution is based on the mode $\hat{\beta}=(\hat{\beta}_1,\ldots,\hat{\beta}_J)$ and inverse of the information matrix $V_{\hat{\beta}}=[-H_{\hat{\beta}}]^{-1}$ of $\log\pi^+(\beta|y,b,D)$. If we assume that the β_j's are a priori independent—that is, the matrix B_0^{-1} is block diagonal with the jth block given by B_{0j}^{-1}—then the modal values $\hat{\beta}_j$ can be found in sequence from a few (typically three or four) Newton–Raphson steps with the gradient vector $-B_{0j}(\beta_j-\beta_{0j})+\sum_{i=1}^{n}[y_{ij}-\exp(x_{ij}'\beta_j+b_{ij})]x_{ij}$ and the Hessian matrix $H_{\beta_j}=-B_{0j}-\sum_{i=1}^{n}[\exp(x_{ij}'\beta_j+b_{ij})]x_{ij}x_{ij}'$. In this case, the dispersion matrix $V_{\hat{\beta}}$ of the proposal is also block diagonal with jth block given by $V_{\hat{\beta}_j}=[-H_{\hat{\beta}_j}]^{-1}$. The algorithm is now implemented as follows. We get a candidate value $\beta^*=\{\beta_j^*\}$ from the proposal density $f_T(\beta_j|\hat{\beta}_j,V_{\hat{\beta}_j},\nu)$, a multivariate-$t$ distribution with mean $\hat{\beta}$ and dispersion matrix $V_{\hat{\beta}_j}$, compute the probability of move

$$\begin{aligned}\alpha(\beta,\beta^*|y,b,D)=\min\Bigg\{&\Bigg(\phi_k(\beta^*|\beta_0,B_0^{-1})\\&\times\prod_{j=1}^{J}\{p(y_{\cdot j}|\beta_j^*,b_{\cdot j})f_T(\beta_j|\hat{\beta}_j,V_{\hat{\beta}_j},\nu)\}\Bigg)\\&\Bigg/\Bigg(\phi_k(\beta|\beta_0,B_0^{-1})\prod_{j=1}^{J}\{p(y_{\cdot j}|\beta_j,b_{\cdot j})\\&\times f_T(\beta_j^*|\hat{\beta}_j,V_{\hat{\beta}_j},\nu)\}\Bigg),1\Bigg\},\end{aligned}\tag{8}$$

and accept the candidate value β^* with probability $\alpha(\beta,\beta^*|y,b,D)$.

If the dimension of β is large, it may turn out that sampling the whole β vector in one block (as in the preceding approach) may produce many rejections. In an effort to deal with such cases, one may revise the (vector) components of β—namely, the β_j's—one at a time. In this alternative approach, one uses the same ingredients as in the preceding method except that now β_j is revised through a sequence of M–H steps. The probability of move for β_j is easily seen to be

$$\begin{aligned}\alpha(\beta_j,\beta_j^*|y_{\cdot j},b_{\cdot j},D)=\min\Bigg\{&\Bigg(\phi_{k_j}(\beta_j^*|\beta_{0j},B_{0j}^{-1})p(y_{\cdot j}|\beta_j^*,b_{\cdot j})\\&\times f_T(\beta_j|\hat{\beta}_j,V_{\hat{\beta}_j},\nu)\Bigg)\Bigg/\Bigg(\phi_{k_j}(\beta_j|\beta_{0j},B_{0j}^{-1})\\&\times p(y_{\cdot j}|\beta_j,b_{\cdot j})f_T(\beta_j^*|\hat{\beta}_j,V_{\hat{\beta}_j},\nu)\Bigg),1\Bigg\}.\end{aligned}$$

In practical experimentation we have found that this modification is more efficient than the one-block approach when the dimension of β is large.

2.5 Sampling D^{-1}

Finally, we sample D^{-1} from the density proportional to

$$f_W(D^{-1}|v_0,R_0^{-1})\prod_{i=1}^{n}\phi_J(b_i|0,D).$$

On combining terms, one sees that this density is Wishart,

$$D^{-1}|b \sim \text{Wishart}\left(n+v_0, \left[R_0^{-1}+\sum_{i=1}^{n}(b_i b_i')\right]^{-1}\right), \qquad (9)$$

with degrees of freedom $n+v_0$ and scale matrix $[R_0^{-1}+\sum_{i=1}^{n}(b_i b_i')]^{-1}$.

2.6 Extensions

The basic model just presented can be generalized by letting $b_i|D, u_i \sim \text{N}_J(0, D/u_i)$, where $u_i \sim \text{gamma}(\gamma_0/2, \gamma_0/2)$ and γ_0 is a prespecified hyperparameter. Integration over u_i leads to a multivariate-t distribution for b_i with γ_0 df, $b_i|D \sim \text{MVT}(0, D, \gamma_0)$, and the marginal distribution of y_i is obtained as before as a J-dimensional integral over the joint distribution of y_i and b_i.

This model allows for thicker tails of the distribution for the latent effects relative to the Poisson lognormal model. Although closed-form expressions for the marginal moments $\text{E}(y_i)$ and $\text{var}(y_i)$ are not available, the computation of the posterior distribution requires only two relatively minor modifications to the MCMC algorithm. The first change occurs in the simulation of the full conditional distribution of b_i, where the conditional Poisson density of cluster i is now multiplied by a multivariate-t distribution rather than by a multivariate-normal so that the ith target density should be written as

$$\pi(b_i|y_i, \beta, D) = c_i f_T(b_i|0, D, \gamma_0) \prod_{j=1}^{J} \exp[-\exp(x_{ij}'\beta_j + b_{ij})] \times [\exp(x_{ij}'\beta_j + b_{ij})]^{y_{ij}}.$$

Second, the distribution of D, conditional on b and u, is again of the Wishart form

$$D^{-1}|b, u \sim \text{Wishart}\left(n+v_0, \left[R_0^{-1}+\sum_{i=1}^{n}(\bar{b}_i \bar{b}_i')\right]^{-1}\right),$$

where $\bar{b}_i = b_i\sqrt{u_i}$. Finally, to obtain values for u_i, one can directly draw from the full conditional distribution $u_i|D, b_i \sim \text{gamma}(v/2, w/2)$, where $v = \gamma_0 + J$ and $w = \gamma_0 + b_i' D^{-1} b_i$.

Another possible extension is to allow the mean of b_i to depend on a given set of covariates W_i to model correlation between those covariates and the latent effects. With this in mind, one could let $b_i|\delta, D \sim \text{N}_J(W_i\delta, D)$. The required changes to the MCMC algorithm are minor. The sampling of β is unaffected by this change, while the full conditional of b_i conditioned on δ has the same form as in (5) except that the prior of b_i now includes a nonzero mean $W_i\delta$. Finally, the sampling of δ and D given b_i follows from standard updates for normal models. For example, under a normal prior on δ, the full conditional of δ given $\{b_i\}$ and D is normal, while that of D^{-1} given $\{b_i\}$ and δ is Wishart as given in (9) with the scale matrix modified to incorporate the nonzero mean of b_i.

3. APPLICATIONS

We illustrate the use of the proposed algorithm on two different datasets. In the first application, we jointly model six measures of medical-care demand by the elderly. The second application is concerned with a high-dimensional problem on the number of airline incidents recorded for 16 U.S. passenger air carriers between 1957 and 1986. In each application, our algorithm is run for 6,000 iterations following a burn-in phase of 500 iterations. The results are robust to the starting values of β [which was the maximum likelihood estimator (MLE) from independent Poisson regressions] and D (which was .1 times the identity matrix). In effect, the algorithm required no user intervention beyond the specification of the model and prior hyperparameters.

3.1 Health-Care Utilization

Deb and Trivedi (1997) estimated independent count-data models for six measures of medical-care demand by the elderly using a sample from the 1987 National Medical Expenditure Survey. One question of substantive interest is the extent to which the use of health services depends on insurance coverage. The six measures are the number of visits to a physician in an office setting (OFP), the number of visits to a nonphysician in an office setting (OFNP), the number of visits to a physician in a hospital outpatient setting (OPP), the number of visits to a nonphysician in a hospital outpatient setting (OPNP), the number of visits to an emergency room (EMR), and the number of hospital stays (HOSP). The sample comprises 4,406 observations on individuals aged 66 or over. For each of these individuals, the sample contains a total of 16 explanatory variables on insurance coverage, health status, and other socioeconomic and demographic characteristics.

Deb and Trivedi treated the six counts as independent and applied univariate finite mixture models to each count. Munkin and Trivedi (1999, henceforth MT) used the same data to estimate a bivariate model for EMR and HOSP by simulated maximum likelihood. Instead of confining ourselves to two counts, we consider the joint determination of all six counts using the model specified previously. This extension should yield a superior model of the demand for health care because the different demand components are likely to be correlated, since any omitted factor in one equation, such as certain aspects of health, will influence other components as well. Moreover, it would be unwise to impose equicorrelation from the outset. Some components could be more closely related than others, and negative correlations may be possible as well if, for instance, different types of health provisions are substitutes.

Since there are 17 regressors in each of the six equations, we use the version of the MCMC algorithm in which the β_j are drawn consecutively for the six equations. The priors for the estimation are defined by the hyperparameters $\beta_0 = 0, B_0^{-1} = 0.01I_7; \nu_0 = 12, R_0 = I_6$, to reflect weak prior information.

Table 1 contains the prior-posterior summary for the regression coefficients. To save space, we only display the results for the emergency-room-visit and hospital-stay equations. The results for this part of the model are directly comparable to those of MT, although, as stated previously, they considered a bivariate model, whereas in our case the two equations are estimated jointly with the four other equations in a full six-variate model for all available health-demand variables.

Table 1. Posterior Summary for β_5 and β_6 Based on the MCMC Simulation Output From Six-Variate Model

Variable	Mean	Std. dev.	Lower	Upper	Ineff.
Response: EMR					
Excellent health	−0.655	0.207	−1.068	−0.256	3.106
Poor health	0.543	0.115	0.324	0.763	9.383
Chronic condition	0.269	0.030	0.211	0.327	8.220
Limits to activities	0.428	0.101	0.223	0.627	8.846
Northeastern U.S.	0.058	0.113	−0.163	0.280	5.714
Midwestern U.S.	0.045	0.104	−0.158	0.253	6.431
Western U.S.	0.168	0.117	−0.068	0.395	7.931
Age$\times 10^{-1}$	0.132	0.065	0.003	0.256	4.915
Black	0.202	0.129	−0.057	0.454	6.448
Male	0.101	0.089	−0.071	0.276	5.882
Married	−0.129	0.095	−0.317	0.057	6.164
Years of schooling	−0.012	0.012	−0.035	0.011	6.083
Income	−0.001	0.015	−0.031	0.028	5.703
Employed	0.205	0.144	−0.081	0.492	5.018
Private insurance	0.062	0.112	−0.155	0.284	6.034
Medicaid	0.207	0.144	−0.080	0.490	6.713
Constant	−3.843	0.538	−4.923	−2.747	5.927
Response: HOSP					
Excellent health	−0.723	0.210	−1.152	−0.322	3.450
Poor health	0.597	0.110	0.381	0.820	11.538
Chronic condition	0.330	0.028	0.273	0.385	9.016
Limits to activities	0.359	0.097	0.166	0.555	10.278
Northeastern U.S.	0.052	0.111	−0.168	0.266	6.816
Midwestern U.S.	0.184	0.100	−0.010	0.377	7.596
Western U.S.	0.156	0.113	−0.068	0.376	8.533
Age$\times 10^{-1}$	0.223	0.063	0.098	0.348	7.353
Black	0.156	0.127	−0.094	0.404	7.753
Male	0.201	0.088	0.031	0.371	6.630
Married	−0.023	0.093	−0.207	0.158	7.130
Years of schooling	0.011	0.012	−0.012	0.035	7.035
Income	−0.002	0.014	−0.030	0.025	6.291
Employed	0.053	0.145	−0.232	0.341	7.806
Private insurance	0.215	0.112	−0.005	0.431	6.645
Medicaid	0.253	0.144	−0.028	0.536	7.107
Constant	−4.993	0.526	−6.020	−3.933	8.063

NOTE: The prior distributions of all parameters have mean 0 and standard deviation 10. In the table, "Lower" and "Upper" denote the 2.5th percentile and the 97.5th percentile of the simulated draws, respectively, and Ineff. denotes the inefficiency factor. The results are based on 6,500 draws of which the first 500 are discarded.

The table gives several summary measures of the posterior distribution. In addition to the posterior mean and standard deviation, we also display the 2.5th and the 97.5th percentile of the marginal posterior distribution and the inefficiency factor (INEFF) (also called the autocorrelation time) in the estimation of the posterior mean of β. If we let G denote the Monte Carlo sample size, then the inefficiency factor is defined as $1+2\sum_{k=1}^{\infty}\rho(k)$, where $\rho(k)$ is the autocorrelation at lag k for the parameter of interest and the terms in the summation are cut off according to (say) the Parzen window. The inefficiency factors are reasonably small, indicating that the sampler is mixing well.

The table shows that many of the included variables contribute significantly to the observed interpersonal variation in visits. The posterior means are very similar to the simulated MLE's reported by MT. For instance, persons who self-assess their health status as poor are more likely, and those who self-assess their health status as excellent are less likely, to visit an emergency room or stay at a hospital than others. In addition to these two measures of health status, conditions that limit activities of daily living have the most significant and strongest effect on the outcome variables. Gender is significant only in the hospital equation, with a higher risk of hospitalization for men. Similarly, private insurance does not affect the number of emergency-room visits but it increases the expected number of hospital stays.

Table 2. Means and Standard Deviations of the Posterior Distribution of the Correlation Matrix of the Latent Effects

	OFP	OFNP	OPP	OPNP	EMR
OFNP	0.330				
	(0.024)				
OPP	0.164	0.115			
	(0.027)	(0.032)			
OPNP	0.403	0.196	0.478		
	(0.027)	(0.033)	(0.031)		
EMR	0.328	0.113	0.300	0.302	
	(0.034)	(0.040)	(0.038)	(0.045)	
HOSP	0.449	0.081	0.433	0.417	0.945
	(0.031)	(0.039)	(0.037)	(0.038)	(0.012)

Next, in Table 2 we summarize the evidence on the correlation structure for our six-variate model. For this purpose, we computed $C=(\mathrm{diag}(D))^{-1/2}D(\mathrm{diag}(D))^{-1/2}$ for each draw from the posterior sample on D. This gives us a sample from the posterior of C, which we have summarized in Table 2 in terms of the means and standard deviations.

There is a positive correlation between each of the latent effects in the six equations but the correlation structure is not homogeneous. As expected, the equations for emergency-room admittance and hospitalization are closely related. We obtain a mean value for the correlation coefficient of 0.945, which is close to the 0.92 reported by MT. The variances of the latent effects in these two equations, not shown here, have posterior means of 1.70 and 1.77, respectively, which are close to the values reported by MT. Since the latent effects in equations 5 and 6 have similar variances and are highly correlated, it is likely that the correlation structure between the last two equations could be captured by a one-factor model as well, as was also concluded by MT. However, a closer look at Table 2 shows that such a conclusion does not generalize to the other four equations and that a flexible model with a full set of correlated latent effects is needed to adequately describe the correlation structure. One observation is that the pairs of equations representing "serious" and "less serious" health problems are relatively unrelated. For example, the correlation between visits to nonphysicians in an office situation and hospitalization is just 0.081. Although negative correlations are not observed in this example, the need for letting the matrix D be unrestricted is evident.

Our discussion so far has concentrated on the regression coefficients and the correlation structure. In addition, the model can be used to obtain predictive distributions. The probability function of the model depends on β, the latent variable b_{ij}, and the covariates x_{ij}. One can compute the average predicted probability of outcomes in equation j, $j=1,\ldots,6$ by integrating $f(y_j|\beta_j, b_{ij}, x_{ij})$, $y_j=0,1,\ldots,$ over the joint posterior distribution of β_j and b and over the observed data distribution of x. In practice, this approach is simple to implement because it requires only the output from the MCMC algorithm.

For each value of β_j^r and $\{b_i^r\}$ from the MCMC simulation, we compute $\theta_{ij}^r = \exp(x_{ij}'\beta_j^r + b_{ij}^r)$. To predict marginal probabilities, we compute

$$\hat{p}_{kj} = \frac{1}{N}\frac{1}{R}\sum_{r=1}^{R}\sum_{i=1}^{N} f(k_j|\theta_{ij}^r), \qquad k_j = 0, 1, \ldots, \tag{10}$$

where f is the Poisson probability function and R is the number of iterations. The prediction of joint probabilities is, however, more interesting because in that case the proper modeling of the correlation structure is important. Predictions of joint probabilities within the context of our model are simple since, conditional on the latent effects, the six equations have independent Poisson distributions and the joint probabilities can thus be obtained through multiplication. Thus, all one needs to do in (10) is to replace $f(k_j|\theta_{ij}^r)$ by $\prod_{j=1}^{6} f(k_j|\theta_{ij}^r)$. Finally, we can construct out-of-sample predictions. In this case, $\{b_i\}$ for the new data points is sampled from a normal distribution conditioned on each sampled value of D.

The predictions from the model can be used to design a heuristic diagnostic check of model fit. The high dimensionality of the joint distribution means that one can consider model fit at various levels. One possibility is to focus on the marginal distributions of the observed data. Table 3 makes such a comparison for the outcomes 0, 1, 2, 3, 4, and 5 or greater. As expected, the in-sample predictions tend to be better than the out-of-sample predictions. With the exception of the first equation, the empirical frequency distributions of the data are traced closely by the predictions. In the case of the OFP response, the high frequency of zeros for the given

Table 3. In- and Out-of-sample Predictions of Marginal Probabilities

		$P(y_j=0)$	$P(y_j=1)$	$P(y_j=2)$	$P(y_j=3)$	$P(y_j=4)$	$P(y_j>4)$
				In-Sample			
OFP	predicted	0.116	0.137	0.123	0.103	0.085	0.436
	empirical	0.155	0.109	0.097	0.095	0.087	0.457
OFNP	predicted	0.674	0.134	0.055	0.030	0.020	0.087
	empirical	0.682	0.129	0.047	0.028	0.024	0.090
OPP	predicted	0.771	0.122	0.042	0.020	0.011	0.034
	empirical	0.771	0.119	0.046	0.017	0.012	0.035
OPNP	predicted	0.842	0.092	0.028	0.012	0.007	0.020
	empirical	0.838	0.098	0.028	0.011	0.006	0.019
EMR	predicted	0.825	0.125	0.031	0.011	0.004	0.004
	empirical	0.818	0.133	0.031	0.012	0.002	0.004
HOSP	predicted	0.810	0.131	0.035	0.013	0.005	0.006
	empirical	0.804	0.136	0.040	0.011	0.005	0.004
				Out-of-sample			
OFP	predicted	0.119	0.144	0.128	0.104	0.082	0.423
	empirical	0.159	0.102	0.103	0.093	0.082	0.461
OFNP	predicted	0.678	0.140	0.055	0.029	0.018	0.080
	empirical	0.688	0.124	0.044	0.030	0.021	0.093
OPP	predicted	0.771	0.118	0.040	0.019	0.011	0.041
	empirical	0.776	0.110	0.046	0.021	0.013	0.034
OPNP	predicted	0.848	0.088	0.026	0.011	0.006	0.020
	empirical	0.834	0.097	0.027	0.011	0.009	0.022
EMR	predicted	0.824	0.120	0.030	0.011	0.006	0.009
	empirical	0.821	0.127	0.034	0.013	0.002	0.004
HOSP	predicted	0.819	0.120	0.031	0.013	0.006	0.011
	empirical	0.800	0.140	0.040	0.008	0.005	0.006

NOTE: For out-of-sample predictions the sample was split in half, and the results using the first 2,203 observations were used to predict the distribution of the other half.

Table 4. In- and Out-of-Sample Predictions of Joint Probabilities

	In-sample			*Out-of-sample*		
Event	*Actual*	*Joint*	*Independent*	*Actual*	*Joint*	*Independent*
(0, 0, 0, 0, 0, 0)	0.099	0.067	0.034	0.104	0.068	0.036
(1+, 0, 0, 0, 0, 0)	0.285	0.296	0.258	0.284	0.308	0.264
(0, 1+, 0, 0, 0, 0)	0.014	0.014	0.016	0.014	0.015	0.017
(0, 0, 1+, 0, 0, 0)	0.010	0.010	0.010	0.008	0.010	0.011
(0, 0, 0, 1+, 0, 0)	0.001	0.003	0.006	0	0.003	0.006
(0, 0, 0, 0, 1+, 0)	0.005	0.005	0.007	0.005	0.005	0.008
(0, 0, 0, 0, 0, 1+)	0.002	0.000	0.008	0.002	0.004	0.008

NOTE: The first column gives the values for the sixtuplet under consideration (OFP, OFNP, OPP, OPNP, EMR, HOSP).

mean suggests that alternative models such as the hurdle or finite mixture models (see Deb and Trivedi 1997) might be appropriate.

As stated previously, the real advantage of a joint model lies in the prediction of joint probabilities. Table 4 gives examples for some interesting cases. For instance, the first row gives the joint probability that a person did not use any of the six health services over the sample period. In the sample, this applies to 10% of all individuals. The joint model predicts the proportion of "nonusers" to be 7%. Although this underpredicts the actual outcome, the joint model offers a substantial improvement over the independece model, which underpredicts the nonuse event even more seriously. The next six rows of Table 4 give the joint probabilities of a count of 1 or greater for one of the six types of health services at a time and zero counts for the other five types. In all instances, the predictions of the joint model are far superior to the predictions under independence, both in and out of sample, and the predictions from the joint sample are accurate to the third digit in a number of cases.

3.2 Number of Airline Incidents

In our second application, we reanalyze the airline accident data of Rose (1990) with a view to illustrating an application involving high-dimensional correlated counts. Our sample data consists of 16 U.S. passenger air carriers (from a total of 35) who had complete observations between 1957 and 1986. The dependent variables are the 16 counts of the number of accidents per carrier per year, where accidents are defined as any operation-related occurrence that leads to personal injury or death or substantial damage to the aircraft. Over the sample the number of accidents ranges from 0 to 14 with mean 1.7 and variance 4.9. Similar data were also analyzed by Dionne, Gagné, Gagnon, and Vanasse (1997) for Canadian air accidents during 1976 and 1987 for more than 120 carriers.

Contemporaneous correlation in accident outcomes may be the consequence of omitted industrywide variables—for instance, safety standards set by the Federal Aviation Authority. Such factors can affect all airlines equally, but this need not be the case. In our model, we capture the common effects through airline fixed effects, whereas the additional correlated latent effects allow both for carrier-specific overdispersion and for additional contemporaneous correlation between carriers. Note that, since we condition on fixed effects, the assumption

Table 5. Posterior Summary From the 16 Variate Multivariate-t Count Model Fit to Airlines Data

	Mean	Std. dev.	Median	Lower	Upper	Ineff.
Operating margin	−0.7565	0.9473	−0.7429	−2.5164	1.0512	8.0446
Stage length	−0.0223	0.5035	−0.0184	−1.0292	0.9424	8.0218
Experience	−0.0081	0.0546	−0.0075	−0.1143	0.0988	7.8574
International	0.4556	0.8551	0.4719	−1.1662	2.1957	6.7933

NOTE: The estimated model also includes 29 year and 16 firm-fixed-effect covariates not shown here. "Lower" and "Upper" denote the 2.5th percentile and the 97.5th percentile, respectively, and Ineff. denotes the inefficiency factor. The results are based on 6,500 draws, of which the first 500 are discarded.

that the latent effects are independent of the regressors is not very strong.

Let y_{it} denote the number of accidents for carrier i in year t. By assumption, $y_{it}|\theta_{it}$ is independently Poisson distributed, where $\theta_{it} = d_{it}\exp(\alpha_i + \delta_t + x_{it}'\beta + b_{it})$. The expected number of accidents is assumed to be proportional to the total number of departures d_{it} (in thousands). The covariates include 16 airline fixed effects, 29 year fixed effects, operating margin as a measure of profitability of the airline (OPMARG), average stage length in thousands of miles (AVSTAGE), cumulative airline operating experience in billions of aircraft miles (EXPER), and the fraction of total departures that are international flights (INTL) (see Rose 1990 for further details).

In this setup, we allow for airline-specific overdispersion and additional contemporaneous correlations between the accident rates of the 16 carriers by assuming that $b_t = (b_{i1}, \ldots, b_{i16})$ is distributed as multivariate-t with mean vector equal to 0, dispersion matrix D, and 10 df. We employ the following hyperparameters: $\beta_0 = 0$, $B_0^{-1} = 0.01I_7$; $\nu_0 = 32$, $R_0 = I_{16}$, $\gamma_0 = 10$, which imply that the prior mean on the diagonal elements of D is approximately $1/32 = .03$ (indicating small heterogeneity) but with fairly large prior variance (due to the low value of the degree of freedom). Our MCMC algorithm for this model, discussed previously, is run for 6,000 iterations beyond a burn-in sample of 500 cycles.

In Table 5 we summarize the posterior distribution of some of the elements of β. After controlling for the airline fixed effects, the year effects, and the correlation among the outcomes, the interpercentile ranges of the marginal posterior distributions of each of the four airline-specific covariates include 0. It is rather more difficult to summarize the posterior distribution of D given that D contains 136 parameters. To give some idea of the posterior distribution, we report in Figure 1 the posterior boxplots of the 16 diagonal elements of D along with the autocorrelation plots of $D_{1,1}$, $D_{9,9}$, $D_{12,12}$, and $D_{16,16}$. We see that the posterior distributions of the diagonal elements are quite similar with median values ranging from about 0.06 to 0.07. Naturally, the posterior distributions are skewed because they are bounded from below at 0. Note that the autocorrelations in the sampled output decline quickly, indicating that the sampler is mixing well. Substantively, in contrast to the previous example, in which the pattern of D was complex, the evidence here points to a correlation matrix that neither indicates airline-specific heteroscedasticity nor substantial contemporaneous correlations (the off-diagonal elements are not displayed; they tend to be close to 0). Overall, this example provides another illustration of the efficacy of our method in high-dimensional count-data models that (as far

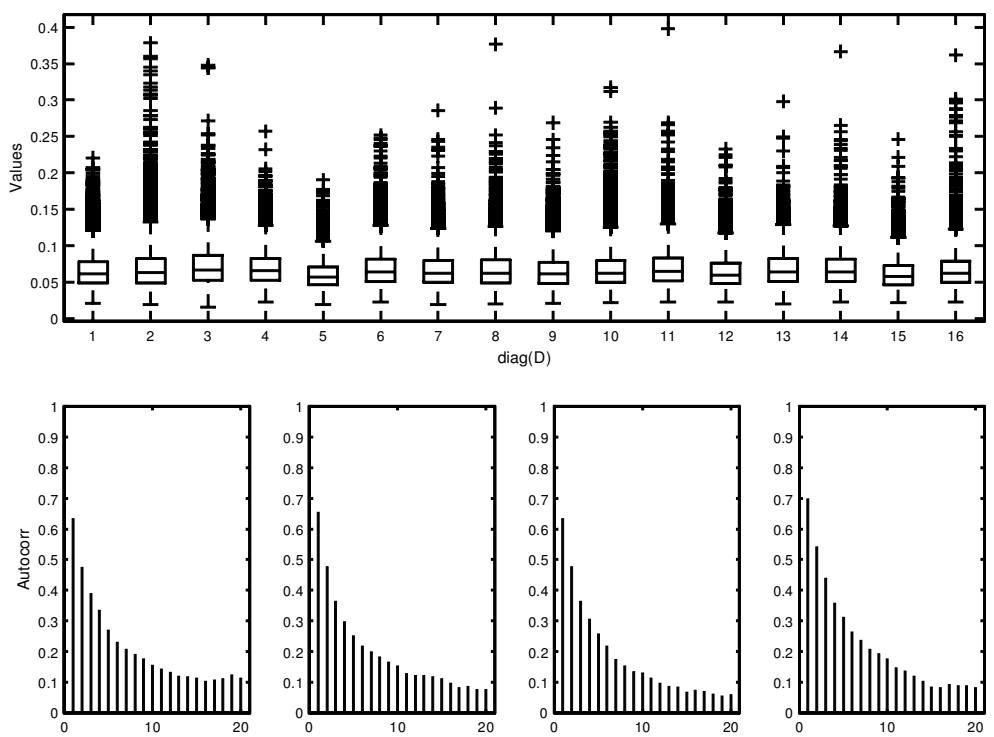

Figure 1. Posterior Boxplots of Diag(D) and Autocorrelation Functions of $D_{1,1}$, $D_{9,9}$, $D_{12,12}$, and $D_{16,16}$ in Airline Count-Data Example.

as we are aware) have never before been fit with this level of generality on the correlation structure.

4. CONCLUDING REMARKS

This article develops a general simulation-based approach for the analysis of multivariate count data. In our model, the correlation among the counts is modeled by assuming that the counts are independent Poisson variates, conditioned on a vector of correlated latent effects. Correlation among the counts is achieved by letting the latent effects be distributed as multivariate log normal or multivariate log-t. The correlation structure of the random effects is taken to be fully general. We develop an MCMC-based approach to estimate the model and show that the method is practical even in high-dimensional problems.The method is applied in the joint analysis of six alternative measures of health-care usage, as well as in the analysis of a panel dataset on air-traffic incidents. In both cases, interesting information on the underlying correlation structure of the counts is recovered.

[Received July 1999. Revised January 2001.]

REFERENCES

Aitchison, J., and Ho, C. H. (1989), "The Multivariate Poisson-log Normal Distribution," *Biometrika*, 76, 643–653.

Blundell, R., Griffith, R., and Van Reenen, J. (1995), "Dynamic Count Data Models of Technological Innovation," *Economic Journal*, 104, 333–344.

Cameron, C., and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge, U.K.: Cambridge University Press.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis–Hastings Algorithm," *The American Statistician*, 49, 327–335.

——— (1996), "Markov Chain Monte Carlo Simulation Methods in Econometrics," *Econometric Theory*, 12, 409–431.

Chib, S., Greenberg, E., and Winkelmann, R. (1998), "Posterior Simulation and Bayes Factors in Panel Count Data Models," *Journal of Econometrics*, 86, 33–54.

Deb, P., and Trivedi, P. K. (1997), "Demand for Medical Care by the Elderly: A Finite Mixture Approach," *Journal of Applied Econometrics*, 12, 313–336.

Dionne, G., Gagné, R., Gagnon, F., and Vanasse, C. (1997), "Debt, Moral-hazard and Airline Safety: An Empirical Evidence," *Journal of Econometrics*, 79, 379–402.

Gurmu, S., and Elder, J. (1998), "Estimation of Multivariate Count Regression Models With Applications to Health Care Utilization," mimeo, University of Virginia, Dept. of Economics.

Hausman, J., Hall, B. H., and Griliches, Z. (1984), "Econometric Models for Count Data With an Application to the Patents–R&D Relationship," *Econometrica*, 52, 909–938.

King, G. (1989), "A Seemingly Unrelated Poisson Regression Model," *Sociological Methods & Research*, 17, 235–255.

Jung, R. C., and Winkelmann, R. (1993), "Two Aspects of Labor Mobility: A Bivariate Poisson Regression Approach," *Empirical Economics*, 18, 543–556.

Munkin, M. K., and Trivedi, P. K. (1999), "Simulated Maximum Likelihood Estimation of Multivariate Mixed-Poisson Regression Models, With Application," *Econometrics Journal*, 2, 29–48.

Rose, N. L. (1990), "Profitability and Product Quality: Economic Determinants of Airline Safety Performance," *Journal of Political Economy*, 98, 944–964.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.

Winkelmann, R. (2000), *Econometric Analysis of Count Data* (3rd ed.) Heidelberg: Springer-Verlag.

Wooldridge, J. M. (1997), "Multiplicative Panel Data Models Without the Strict Exogeneity Assumption," *Econometric Theory*, 13, 667–678.