Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity

Eun Sug Park
Associate Research Scientist
Texas Transportation Institute, 3135 TAMU
College Station, TX 77843-3135
phone: 979/845-9942, fax: 979/845-6481
email: e-park@tamu.edu

Dominique Lord
Assistant Professor
Department of Civil Engineering
Texas A&M University
Texas Transportation Institute, 3135 TAMU
College Station, TX 77843-3135
Phone : (979) 458-1218
E-mail : d-lord@ttimail.tamu.edu

Prepared for presentation for
86th Annual Meeting of the Transportation Research Board, Washington, D.C.

Words: 4,290 + 4 tables = 5,290 words
Submitted: July 2006
Revised: November 2006

ABSTRACT

This paper introduces a new multivariate approach for jointly modeling crash counts by severity data based on Multivariate Poisson-Lognormal models. Although the crash frequency by severity data are multivariate in nature, they have often been analyzed by modeling each severity level separately without taking into account correlations that exist among different severity levels. The new Multivariate Poisson-Lognormal regression approach can cope with both over-dispersion and a fully general correlation structure in the data as opposed to the recently suggested Multivariate Poisson regression approach that allows for neither over-dispersion nor a general correlation structure in the data. The new method is applied to the multivariate crash counts obtained from the intersections in California for 10 years. The results show promise towards the goal of obtaining more accurate estimates by accounting for correlations in the multivariate crash counts and over-dispersion.

INTRODUCTION

There has been considerable research on crash data analysis and statistical modeling (1-12). Crash data are often collected in terms of crash frequencies and severity levels. Examples of severity levels are fatal (K), incapacitating-injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO or O). Although the crash frequency by severity data are multivariate in nature, they have often been analyzed by modeling each severity level separately without taking into account correlations that exist among different severity levels. Usually, statistical models are produced for all crash severity all together (i.e., often referred to as KABCO) or for different crash severity levels, such as fatal and non-fatal crashes (e.g., KAB) or PDO crashes (e.g., O).

Treating the correlated crash counts as independent and applying a univariate model to each count leads to less precise estimates for the effects of factors on crash risk. Unfortunately, there has not been much research on jointly modeling crash counts of different severity levels in highway safety. Notable exceptions include articles by Tunaru (13), Bijleveld (14), Miaou and Song (12), Song et al. (15), and Ma and Kockelman (16). Ma and Kockelman (16) adapted a Multivariate Poisson (MVP) regression approach developed by Tsionas (17) to assess the effects of various covariates on the multivariate crash counts by severity. The MVP regression models, however, do not allow for over-dispersion that is often observed in the crash data. In addition, the MVP regression models used by Ma and Kockelman (16) assume that the covariances for different severity levels are all identical (although the assumption of equal covariances has been relaxed in an extended multivariate Poisson regression model developed by Karlis and Meligkotsidou (18)) and non-negative, which is very restrictive. It is possible that different severity levels may have different covariances, and also possibility of negative correlations cannot be entirely excluded.

A Multivariate Poisson-Lognormal (MVPLN) regression approach developed by Chib and Winkelmann (19) can serve as a good alternative to a pure MVP regression approach for analysis of multivariate crash count data because the MVPLN can account for over-dispersion and a fully general correlation structure while the MVP cannot. Also, MVPLN regression models are more general than Multivariate Negative Binomial regression models in the sense that the former can account for negative correlations while the latter cannot. Although these models have already been developed in statistics (see, e.g., 20), there have been almost no attempts to employ those models in roadway safety to model multivariate crash frequency by severity data. One exception is Tunaru (13) who introduced a Multivariate Poisson-Lognormal model in the non-regression context (ranking the sites with accidents) to take into account general correlation structures. However, he did not consider any covariates.

Implementation of MVPLN models is not straightforward. It needs to be noted that none of the existing statistical software have the ability to estimate these models as built-in functions. As mentioned in Chib and Winkelmann (19), it is necessary to adapt simulation-based methods such as a Markov chain Monte Carlo (MCMC) simulation method (see, e.g., 21-23) to cope with the multiple integral in the likelihood function. To estimate MVPLN models, the MATLAB (24) codes tailored to multivariate crash data modeling have been developed according to the MCMC algorithm of Chib and Winkelmann (19).

This paper presents the analysis of multivariate crash count data by severity using the MVPLN regression models implemented by MCMC to assess the effects of covariates. The

models were developed using crash data collected at three-legged unsignalized intersections in California.

MULTIVARIATE POISSON-LOGNORMAL MODELS

The underlying model and the implementation algorithm based on which the MCMC codes were developed are re-described here in the context of the crash count data. Mathematical details can be found in Chib and Winkelmann (19). Because the dimensions of the matrices/vectors given in Chib and Winkelmann (19) were sometimes inconsistent, they are redefined here for clarification purposes.

Modeling Framework

Let **Y** denote an *n* by *J* matrix of the multivariate crash counts where *n* corresponds to the number of intersections and *J* corresponds to the number of different severity types. Let **b** denote an *n* by *J* matrix of latent effects of which rows, $b_i = (b_{i1}, \ldots, b_{iJ})$, $i = 1,\ldots,n$, correspond to a set of *J* intersection and outcome-specific latent effects. Let *k* be the number of covariates and let *X* denote an *n* by *k* matrix of covariates of which rows, $x_i = \left(x_{i1}, \quad \cdots, \quad x_{ik}\right)$, are the *k*-dimensional row vectors corresponding to the *i*th intersection as follows:

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Let $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 & \cdots & \beta_J \end{bmatrix}$ denote a *k* by *J* matrix of the regression coefficients of which columns, $\beta_j = \begin{pmatrix} \beta_{1j} \\ \vdots \\ \beta_{kj} \end{pmatrix}$, are the *k*-dimensional column vectors consisting of parameters for the crash count of *j*th severity type. Suppose that, conditional on b_{ij} and parameters $\beta_j \in \mathrm{R}^k$, the crash count of the *j*th severity type at the *i*th intersection, y_{ij}, follows a Poisson distribution with mean $\mu_{ij} = \exp\left(x_i \beta_j + b_{ij}\right)$, i.e.,

$$y_{ij} \left| b_i, \beta_j \right. \sim Poisson\left(\mu_{ij}\right) \tag{1}$$

where

$$\mu_{ij} = \exp\left(x_i \beta_j + b_{ij}\right) \tag{2}$$

for $j=1,\ldots J$ and $i = 1,\ldots,n$. The y_{ij}'s are independent given the μ_{ij}'s.

To model the correlations among the crash counts of *J* different severity types at an intersection, let

$$b_i \,|\, \Sigma \sim N_J\left(0, \Sigma\right), \qquad i = 1,\ldots,n, \tag{3}$$

where Σ is an unrestricted covariance matrix, and N_J denotes J-dimensional multivariate normal distribution. It was shown in Chib and Winkelmann (19) that the variance of y_{ij} is greater than the mean (allowing for over-dispersion) as long as the diagonal elements of Σ are greater than 0, and the covariance between the counts, y_{ij} and y_{is}, can be positive or negative depending on the sign of the $(j,s)^{th}$ element of Σ. Thus, the correlation structure of the crash counts is unrestricted.

Estimation via MCMC

As noted in Chib and Winkelmann (19), the marginal distribution of the counts $y_i = \left(y_{i1}, \quad y_{i2}, \quad \cdots, \quad y_{iJ}\right)$ cannot be obtained by direct computation because it requires the evaluation of a J-variate integral of the Poisson distribution with respect to the distribution of b_i. The MCMC simulation is thus employed for parameter estimation under a Bayesian framework. For the prior on the parameters, we assume that $\left(\beta_1, \beta_2, \cdots, \beta_J, \Sigma\right)$ independently follow the distributions $\beta_j \sim N_k\left(\beta_0, B_0^{-1}\right)$, $j = 1, \cdots, J$, $\Sigma^{-1} \sim Wishart\left(R_0, r_0\right)$, where $\left(\beta_0, B_0, r_0, R_0\right)$ are known hyperparameters and $Wishart\left(\cdot,\cdot\right)$ is the Wishart distribution (see, e.g., 25) with scale matrix R_0 and degrees of freedom parameter r_0. Then, the joint posterior density is proportional to:

$$\begin{aligned} &\text{Posterior} \propto \text{likelihood} \times \text{prior} \\ &= f_W\left(\Sigma^{-1} | r_0, R_0\right) \prod_{j=1}^{J} \phi_k\left(\beta_j \,|\, \beta_0, B_0^{-1}\right) \prod_{i=1}^{n} \left\{ \prod_{j=1}^{J} f\left(y_{ij} \,|\, \beta_j, b_{ij}\right) \right\} \phi_J\left(b_i \,|\, 0, \Sigma\right), \end{aligned} \tag{4}$$

where f_W is the Wishart density and ϕ_k and ϕ_J are the k-variate and the J-variate normal density, respectively.

We make use of three move types as in Chib and Winkelmann (19) in implementing MCMC: (a) Sampling **b**, (b) Sampling $\boldsymbol{\beta}$, and (c) Sampling Σ^{-1}. We present those three steps briefly again. Some notational errors/typos found in Chib and Winkelmann (19) have been corrected here.

(a) Sampling $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$ where each b_i is a J-dimensional row vector.

The full conditional posterior density for b_i, $\pi\left(b_i \,|\, \cdots\right)$, is not given by any known density (see 19) and requires the Metropolis-Hastings (M-H) algorithm (see, e.g., 26). A multivariate-t distribution with degrees of freedom v_1, the location parameter $\hat{b}_i$ and the scale

parameter $V_{\hat{b}_i}$, $f_T\left(b_i \middle| \hat{b}_i, V_{\hat{b}_i}, \nu_1\right)$, is used as a proposal density for $b_{i\cdot}$. Here, ν_1 is a tuning parameter and $\hat{b}_i$ and $V_{\hat{b}_i}$ are the mode and the inverse of minus the Hessian matrix of $\log \pi^+\left(b_i \middle| y_i, \boldsymbol{\beta}, \Sigma\right)$ at the mode $\hat{b}_i$ where 'log' denotes a natural log and

$$\log \pi^+\left(b_i \middle| y_i, \boldsymbol{\beta}, \Sigma\right) = -0.5\ln\left|2\pi\Sigma\right| - 0.5\left(b_i\Sigma^{-1}b_i'\right) + \sum_{j=1}^{J}\left[-\exp\left(x_i\beta_j + b_{ij}\right) + y_{ij}\left(x_i\beta_j + b_{ij}\right)\right]. \quad (5)$$

To find $\hat{b}_i$ and $V_{\hat{b}_i} = \left(-H_{\hat{b}_i}\right)^{-1}$, the Newton-Raphson algorithm with the gradient vector $g_{b_i} = -b_i\Sigma^{-1} + \left[y_i - \exp\left(x_i\boldsymbol{\beta} + b_i\right)\right]$ and Hessian matrix $H_{b_i} = -\Sigma^{-1} - diag\left\{\exp\left(x_i\boldsymbol{\beta} + b_i\right)\right\}$ can be used. A proposal b_i^* drawn from $f_T\left(b_i \middle| \hat{b}_i, V_{\hat{b}_i}, \nu_1\right)$ is then accepted with probability

$$\min\left\{\frac{\pi^+\left(b_i^* \middle| y_i, \beta, \Sigma\right) f_T\left(b_i \middle| \hat{b}_i, V_{\hat{b}_i}, \nu_1\right)}{\pi^+\left(b_i \middle| y_i, \beta, \Sigma\right) f_T\left(b_i^* \middle| \hat{b}_i, V_{\hat{b}_i}, \nu_1\right)}, 1\right\}.$$

(b) Sampling $\boldsymbol{\beta} = \left[\beta_1 \quad \beta_2 \quad \cdots \quad \beta_J\right]$ where each β_j is a *k*-dimensional vector.

The full conditional posterior density for $\boldsymbol{\beta}$ is not given by any known density either and it requires the Metropolis-Hastings (M-H) algorithm. We use a "block-at-a-time" Metropolis-Hastings algorithm to sample β_j ($j = 1, \cdots, J$) one at a time. A multivariate-*t* distribution with degrees of freedom ν_2, the location parameter $\hat{\beta}_j$ and the scale parameter $V_{\hat{\beta}_j}$, $f_T\left(\beta_j \middle| \hat{\beta}_j, V_{\hat{\beta}_j}, \nu_2\right)$, can be used as a proposal density. Here, ν_2 is a tuning parameter and $\hat{\beta}_j$ and $V_{\hat{\beta}_j}$ are the mode and the inverse of minus the Hessian matrix of $\log \pi^+\left(\beta_j \middle| y_{\cdot j}, b_{\cdot j}\right)$ at the mode $\hat{\beta}_j$ where 'log' denotes a natural log, $y_{\cdot j}$ and $b_{\cdot j}$ denote the *j*th columns of the matrices **Y** and **b**, respectively, and

$$\log \pi^+\left(\beta_j \middle| y_{\cdot j}, b_{\cdot j}\right)$$
$$= -0.5\log\left|2\pi B_{0j}^{-1}\right| - 0.5\left(\left(\beta_j - \beta_{0j}\right)' B_{0j}\left(\beta_j - \beta_{0j}\right)\right) + \sum_{i=1}^{n}\left[-\exp\left(x_i\beta_j + b_{ij}\right) + y_{ij}\left(x_i\beta_j + b_{ij}\right)\right]. \quad (6)$$

To find $\hat{\beta}_j$ and $V_{\hat{\beta}_j} = \left(-H_{\hat{\beta}_j}\right)^{-1}$, the Newton-Raphson algorithm with the gradient vector $-B_{0j}\left(\beta_j - \beta_{0j}\right) + \sum_{i=1}^{n}\left[y_{ij} - \exp\left(x_i\beta_j + b_{ij}\right)\right]x_i'$ and Hessian matrix

$H_{\beta_j} = -B_{0j} - \sum_{i=1}^{n}\left[\exp\left(x_i\beta_j + b_{ij}\right)\right]x_i'x_i$ can be used. A proposal β_j^* drawn from $f_T\left(\beta_j \middle| \hat{\beta}_j, V_{\hat{\beta}_j}, \nu_2\right)$ is then accepted with probability

$$\min\left\{\frac{\pi^+\left(\beta_j^* \middle| y_{\cdot j}, b_{\cdot j}\right) f_T\left(\beta_j \middle| \hat{\beta}_j, V_{\hat{\beta}_j}, \nu_2\right)}{\pi^+\left(\beta_j \middle| y_{\cdot j}, b_{\cdot j}\right) f_T\left(\beta_j^* \middle| \hat{\beta}_j, V_{\hat{\beta}_j}, \nu_2\right)}, 1\right\}.$$

(c) Sampling Σ^{-1}.

The Gibbs sampling algorithm is used to sample Σ^{-1} because the full conditional posterior distribution of Σ^{-1} is given by:

$$\Sigma^{-1}\left|\mathbf{b} \sim Wishart\left(n + \upsilon_0, \left[R_0^{-1} + \sum_{i=1}^{n} b_i' b_i\right]^{-1}\right)\right. . \quad (7)$$

Inferences based on MCMC samples

Thousands of (or millions of if necessary) samples can be simulated indirectly from the joint posterior distribution using the above MCMC algorithm. Here, the samples represent the values of the parameters, and inferences (point estimates, uncertainty estimates, and/or interval intervals) on the parameters can be directly made based on those samples (often called posterior samples). For instance, the sample mean and sample standard deviation of the posterior samples of β can be used as the point estimate and the corresponding uncertainty estimate (standard error) for β. Also, the 2.5th percentile and the 97.5th percentile can be used to construct the 95% credible interval for the elements of β, which is another useful way of representing uncertainty. It needs to be emphasized that convergence of the chain has to be ensured before one makes any inferences.

APPLICATION TO CALIFORNIA INTERSECTION CRASH DATA

The MVPLN models are applied to the crash count data of five different severity levels (Sev1: fatal (K), Sev2: incapacitating-injury (A), Sev3: non-incapacitating injury (B), Sev4: minor injury (C), Sev5: property damage only (PDO or O) collected from 451 three-legged unsignalized intersections in California, obtained through the Highway Safety Information System (HSIS). Although the original data contained the crash counts from 537 intersections, only the intersections having 10 years of crash data history were retained (resulting in 451 intersections). There were 77 fatal injuries, 202 accidents of severity 2, 738 accidents of severity 3, 865 accidents of severity 4, and 2,857 PDO accidents at those 451 intersections for 10 years. Table 1 contains summary statistics of the variables of interest. In the table, the unit of crash frequency is the number of crashes per intersection for 10 years. The major and minor roads of

the intersection are defined as a function of the entering traffic flow. The legs with the highest entering flows are defined as major AADT.

TABLE 1. Summary Statistics of the Variables for California Intersection Data

Variable Name	Mean	Std Dev	Min	Max
Dependent Variables				
Sev1	0.1707	0.5204	0	5
Sev2	0.4479	0.9609	0	6
Sev3	1.6364	2.5159	0	20
Sev4	1.9180	3.5571	0	28
Sev5	6.3348	9.9493	0	88
Independent Variables				
Lighting (1= yes)	0.3525	0.4783	0	1
Painted Left Turn (1= yes)	0.3925	0.4888	0	1
Curb Med Left Turn (1=yes)	0.1330	0.3340	0	1
Rhgt Trn Channel (1=yes)	0.1397	0.3470	0	1
ML Lanes (Nb of Main Lanes)	3.6851	0.7292	2	4
Mountain Terrain (1=yes)	0.1397	0.3470	0	1
Rolling Terrain (1=yes)	0.3570	0.4796	0	1
Logmaj (Logarithm of major AADT)	9.4195	0.7514	7.7956	11.2683
Logmin (Logarithm of minor AADT)	4.9193	1.5148	2.3026	10.0481

Table 2 gives the estimates (posterior means and standard deviations) of the regression coefficients $\boldsymbol{\beta}$ based on a MVPLN model implemented by MCMC using the MATLAB (24) codes specifically developed for this research. Recall that the functional form used for the models was described in Equations 1-3. The dependent variable is defined as the number of crashes per 10 years. To ensure that the chain has converged to the posterior distribution by the end of the burn-in period, trace plots and the autocorrelation function plots of posterior sample values were inspected although those plots are not presented in the paper due to space limitations.

For comparison purposes, we report in Table 2 the estimates obtained by applying the univariate Poisson regression model and the univariate Negative Binomial regression model implemented in SAS (27) as well. It needs to be noted that for an objective comparison the prior distributions of the parameters and the starting values in MCMC implementation have been obtained independently of the SAS results. Here, we use vague priors not requiring much prior knowledge on the parameters to illustrate that the suggested MVPLN models can be applied even without precise prior knowledge. When there exists good prior information on the parameters, however, it can be incorporated by the use of more informative (precise) prior distribution, and it may further improve the precision of the MVPLN models. A good discussion on elicitation of priors in crash data analysis can be found in Schluter et al. (28), for example. Finally, all the variables described in Table 1 were included in the models to facilitate the comparison between the multivariate and univariate models.

TABLE 2. Estimates of Regression Coefficients obtained by Applying the Multivariate Poisson-Lognormal model, the Univariate Poisson regression model, and the Univariate Negative-Binomial regression model

Severity	Variable	Multivariate Poisson-Lognormal model	Univariate Poisson regression	Univariate Negative Binomial regression
Sev1	Constant	**-13.0261** (1.6854)	**-15.2279** (2.0375)	**-14.9638** (2.2200)
	Lighting	-0.5544 (0.3229)	-0.5955 (0.3204)	-0.5704 (0.3470)
	Painted Left Turn	0.5349 (0.2859)	0.5032 (0.2886)	0.5158 (0.3138)
	Curb Med Left Turn	0.4994 (0.3534)	0.6221 (0.3446)	0.6228 (0.3884)
	Rhgt Trn Channel	0.2777 (0.3156)	0.3752 (0.2870)	0.2991 (0.3356)
	ML Lanes	0.2934 (0.2764)	0.2815 (0.3045)	0.2714 (0.3152)
	Mountain	-0.1367 (0.3720)	-0.3431 (0.3764)	-0.1864 (0.4232)
	Rolling	-0.3916 (0.2733)	**-0.5641** (0.2689)	-0.5400 (0.3005)
	Logmaj ADT	**0.8818** (0.1698)	**1.1537** (0.1894)	**1.1188** (0.2088)
	Logmin ADT	**0.2069** (0.0873)	**0.2052** (0.0810)	**0.2223** (0.0921)
				Dispersion: 0.7059
	Pearson Chi-Square/DF		1.2232	1.0667
Sev2	Constant	**-12.5689** (1.2596)	**-13.2302** (1.1873)	**-13.4023** (1.4116)
	Lighting	0.2345 (0.1993)	0.2997 (0.1733)	0.2844 (0.2072)
	Painted Left Turn	**0.5569** (0.2031)	**0.4796** (0.1706)	**0.5572** (0.2023)
	Curb Med Left Turn	0.1780 (0.2856)	0.2229 (0.2431)	0.2290 (0.2882)
	Rhgt Trn Channel	0.2285 (0.2379)	0.3425 (0.1821)	0.2686 (0.2408)
	ML Lanes	0.1625 (0.1737)	0.1571 (0.1563)	0.1490 (0.1703)
	Mountain	0.3866 (0.2667)	0.3106 (0.2294)	0.4187 (0.2762)
	Rolling	**0.4564** (0.1918)	**0.4112** (0.1611)	**0.4710** (0.1910)
	Logmaj	**0.9097** (0.1336)	**1.0435** (0.1186)	**1.0548** (0.1422)
	Logmin	**0.2331** (0.0612)	**0.1899** (0.0492)	**0.1952** (0.0619)
				Dispersion: 0.6070
	Pearson Chi-Square/DF		1.2699	1.0042
Sev3	Constant	**-9.8505** (0.8479)	**-9.9059** (0.5815)	**-10.1854** (0.8482)
	Lighting	0.2081 (0.1360)	**0.2315** (0.0907)	0.2025 (0.1321)
	Painted Left Turn	0.1088 (0.1388)	0.0648 (0.0844)	0.1206 (0.1271)
	Curb Med Left Turn	0.0560 (0.1875)	0.0780 (0.1188)	0.0896 (0.1811)
	Rhgt Trn Channel	0.0793 (0.1619)	**0.2511** (0.1002)	0.0499 (0.1655)
	ML Lanes	0.0417 (0.0995)	0.0404 (0.0692)	0.0491 (0.0911)
	Mountain	**0.4458** (0.1650)	**0.3636** (0.1074)	**0.5708** (0.1691)
	Rolling	0.0734 (0.1329)	0.0447 (0.0846)	0.0885 (0.1257)
	Logmaj	**0.8936** (0.0907)	**0.9463** (0.0604)	**0.9645** (0.0881)
	Logmin	**0.1789** (0.0419)	**0.1608** (0.0262)	**0.1670** (0.0393)
				Dispersion: 0.6048
	Pearson Chi-Square/DF		2.0799	1.0555
Sev4	Constant	**-11.9536** (0.8721)	**-12.4660** (0.5726)	**-11.4316** (0.8863)
	Lighting	**0.5212** (0.1409)	**0.5264** (0.0845)	**0.5422** (0.1394)
	Painted Left Turn	0.0119 (0.1485)	-0.0357 (0.0774)	0.0169 (0.1354)
	Curb Med Left Turn	-0.1958 (0.1990)	-0.1487 (0.1172)	-0.1396 (0.1984)
	Rhgt Trn Channel	0.2490 (0.1789)	**0.2908** (0.0917)	0.3392 (0.1743)
	ML Lanes	0.0134 (0.1007)	0.0140 (0.0649)	0.0093 (0.0966)

	Mountain	**0.4015** (0.1790)	**0.3253** (0.1007)	**0.4683** (0.1837)
	Rolling	0.0518 (0.1451)	0.0569 (0.0787)	0.0536 (0.1353)
	Logmaj	**1.0857** (0.0926)	**1.2034** (0.0593)	**1.0921** (0.0938)
	Logmin	**0.2317** (0.0442)	**0.1982** (0.0240)	**0.1997** (0.0417)
				Dispersion: 0.8015
	Pearson Chi-Square/DF		2.8881	1.2074
Sev5	Constant	**-9.9596** (0.6670)	**-10.1806** (0.3065)	**-9.6546** (0.6358)
	Lighting	**0.4203** (0.1051)	**0.3544** (0.0465)	**0.4881** (0.1049)
	Painted Left Turn	**-0.2159** (0.1127)	**-0.2326** (0.0420)	**-0.2327** (0.1027)
	Curb Med Left Turn	-0.1494 (0.1482)	**-0.1836** (0.0611)	-0.2024 (0.1471)
	Rhgt Trn Channel	0.0715 (0.1263)	**0.1864** (0.0525)	0.1016 (0.1311)
	ML Lanes	0.1257 (0.0723)	**0.1041** (0.0373)	**0.1423** (0.0692)
	Mountain	**0.5337** (0.1347)	**0.5352** (0.0533)	**0.5966** (0.1376)
	Rolling	0.1260 (0.1046)	**0.1403** (0.0437)	0.0699 (0.1004)
	Logmaj	**0.9777** (0.0717)	**1.0593** (0.0315)	**0.9829** (0.0676)
	Logmin	**0.2493** (0.0333)	**0.2193** (0.0132)	**0.2291** (0.0321)
				Dispersion: 0.6225
	Pearson Chi-Square/DF		5.4932	1.1823

Notes: 1. Multivariate Poisson-Lognormal model was implemented by MCMC coded in MATLAB (24).
2. Univariate Poisson regression and Univariate Negative Binomial regression were implemented in SAS (27).
3. Numbers in parentheses represent uncertainty estimates; posterior standard deviations under Multivariate Poisson lognormal model and standard errors under Univariate Poisson regression model and Univariate Negative binomidal regression model, respectively.
4. Significant (at α=0.05) effects are shown in bold.

It can be observed from Table 2 that for Sev1 and Sev2 all three models give similar results in terms of the estimated model coefficients and their significance except for Rolling of Sev1 (which was significant only under the univariate Poisson regression model). For Sev3-Sev5, however, univariate Poisson regression models give significantly different results (in terms of both point estimates and uncertainty estimates) from those of MVPLN models or univariate Negative Binomial regression models. For Sev3-Sev5, it appears that under the univariate Poisson regression model the standard errors are seriously underestimated and as a result many of the covariates are incorrectly declared to be significant. Note that the values of Pearson's Chi-Square divided by degrees of freedom for univariate Poisson regression models are considerably greater than 1 for Sev3-Sev5, which indicates an apparent over-dispersion problem. It is well-known that the more over-dispersion, the more seriously standard errors are underestimated, in which case those standard errors are not correct estimates of true uncertainties and the corresponding interval estimates will not be able to capture the true parameter values. This problem cannot be overcome with MVP regression models either because over-dispersion is not accounted for by those models. It needs to be emphasized that for the unbiased estimates the small standard errors (more precise estimates), only when they are not underestimated, lead to more accurate parameter estimates.

MVPLN models and univariate Negative Binomial models give, in general, consistent results in terms significance of model coefficients. Notice, however, that for Sev1 the uncertainty estimates from the MVPLN model are noticeably smaller than those from the univariate Negative Binomial model. Unlike univariate Poisson regression models, both MVPLN models and univariate Negative Binomial regression models are able to account for over-

dispersion, and their standard errors can serve as good estimates of true uncertainties. This supports that by accounting for correlation in multivariate crash frequency by severity a MVPLN model leads to more precise parameter estimates than a univariate Negative Binomial model does.

With regards to the interpretation of the models' output, Table 2 shows that the coefficients sometimes change from positive to negative values and vice versa for different crash severity levels (e.g., painted left-turn bay). This characteristic can indicate that some variables may have a different effect given the severity outcome of the crash. On the other hand, some coefficients appear to be counterintuitive. For example, the presence of lighting is associated with more crashes for Sev4 and Sev5 (all three models). Given the limited number of observations available in this study, it is possible that confounding factors, such as the location where lighting is used, may explain this outcome. As a reviewer pointed out, speed-related factor such as the speed limit could have been an influencing factor on severity. We regret that the speed limit data were not available for this analysis, which might also have caused confounding of Lighting effect. While the effect of confounding factors such as Speed limit may also exist for other severity levels, the effect could be different for different severity levels unless the correlations among different severity levels are very high. It is expected that Sev5 crash counts will be more closely correlated with Sev4 crash counts than with Sev1 crash counts because Sev4 and Sev5 are more similar in nature. It explains why the effect of Lighting is significant only for Sev4 and Sev5. Obviously, these outcomes need to be further investigated. While we recognize the limitation of this database, we would like to emphasize that the purpose of this paper is to present a general methodology that can account for correlations in the multivariate crash counts and illustrate the method on the real crash counts of different severity types.

Table 3 and Table 4 contain the MCMC estimates of the covariance matrix and correlation matrix of the latent effects (generating the correlation structure in the multivariate crash counts) of the MVPLN model, respectively.

TABLE 3 Posterior means of the covariance matrix (Σ) of the latent effects

	Sev1	**Sev2**	**Sev3**	**Sev4**	**Sev5**
Sev1	0.6592	0.4884	0.4743	0.5487	0.4638
Sev2	0.4884	0.7408	0.5251	0.6054	0.4998
Sev3	0.4743	0.5251	0.7224	0.6595	0.5424
Sev4	0.5487	0.6054	0.6595	0.9357	0.6760
Sev5	0.4638	0.4998	0.5424	0.6760	0.6651

TABLE 4 Posterior means of the correlation matrix of the latent effects

	Sev1	**Sev2**	**Sev3**	**Sev4**	**Sev5**
Sev1	1.0000				
Sev2	0.7035	1.0000			
Sev3	0.6904	0.7203	1.0000		
Sev4	0.7030	0.7297	0.8035	1.0000	
Sev5	0.7043	0.7152	0.7834	0.8575	1.0000

Recall that MVP regression models suggested by other researchers (e.g., 16) are very restrictive in the sense that they assume the covariances for different severity levels are all

identical and non-negative as well as no over-dispersion. On the other hand, the new MVPLN regression models that can be implemented by MCMC allow for a fully general correlation structure as well as over-dispersion in the crash data. From Tables 3 and 4, it can be observed that there is a positive correlation between each of the latent effects in the crash counts of five severity types but the correlations for different severity levels are not identical. Thus, as any statistical models, this correlation needs to be incorporated in the estimation of the model.

There are a few important avenues for further work. First, the predicted values obtained from the MVPLN model can be compared with the values estimated from a univariate Negative Binomial Regression model estimated for all crash severities combined (e.g., KABCO) with the output of an ordered logit crash severity model, as proposed by Miaou et al. (29). Traditional tools and the ones proposed by Oh et al. (10) could be used for this comparison analysis. Second, the stability of the MVPLN models subjected to low sample mean values and small sample size should be investigated along with sensitivity analysis for different hyper-prior specifications; crash data often exhibit these two unique properties. As noted by Lord (30) and Lord and Miranda-Moreno (31), statistical models have the potential to become very unstable when they are estimated using this kind of data. Third, the changes in the signs of the coefficients between different crash severity levels need to be investigated further. It is also desired to investigate if there is any confounding between the variable Lighting and the variables that are not included in the data, which derives the counterintuitive signs for the coefficient for the former variable for Severity 4 and Severity 5 crashes.

SUMMARY AND CONCLUSIONS

This paper presented a new multivariate approach for modeling crash counts by severity data based on Multivariate Poisson-Lognormal models employing MCMC as a computational engine. The method was applied to the multivariate crash counts from 451 intersections in California obtained for 10 years. It turned out that not only there are correlations across severity levels but also the correlations are not identical. Neither of univariate modeling approach nor previously suggested Multivariate Poisson regression approach (16) would have revealed it. Over-dispersion in the data was apparent, which again cannot be handled by an MVP approach.

The new Multivariate Poisson-Lognormal regression approach can cope with both over-dispersion and a fully general correlation structure in the data. It was also observed that for fatal crashes (Sev1) the uncertainty estimates from the Multivariate Poisson-Lognormal model are noticeably smaller than those from the univariate Negative Binomial model, which suggests that by accounting for correlation in the multivariate crash counts a Multivariate Poisson-Lognormal model leads to more precise parameter estimates than a univariate Negative Binomial model does.

ACKNOWLEDGEMENTS

The authors would like to thank Mr. Srinivas Geedipally and Mr. Craig Lyon for assisting in collecting and assembling the HSIS data used in this study. The work presented in this paper was initially carried out as part of the research project NCHRP 17-29 and the authors thank the National Academy of Science (NAS) for funding the study. The opinions in this document

reflect the views of the authors only and do not necessarily reflect the points of view of any other sponsoring or contributing individual or agency.

REFERENCES

1. Abbess, C., D. Jarett, and C.C. Wright. Accidents at Blackspots: estimating the Effectiveness of Remedial Treatment, With Special Reference to the "Regression-to-Mean" Effect. *Traffic Engineering and Control*, Vol. 22, No. 10, 1981, pp. 535-542.

2. Hauer, E., J. C. N. Ng, and J. Lovell, Estimation of Safety at Signalized Intersections. In *Transportation Research Record* 1185, TRB, National Research Council, Washington, D.C., 1988, pp. 48-61.

3. Persaud, B.N. and L. Dzbik, Accident Prediction Models for Freeways. In *Transportation Research Record* 1401, TRB, National Research Council, Washington, D.C., 1993, pp. 55-60.

4. Kulmala, R. Safety at Rural Three- and Four-Arm Junctions: Development and Applications of Accident Prediction Models. VTT Publications 233, Technical Research Centre of Finland, Espoo, 1995.

5. Poch, M., and F.L. Mannering. Negative binomial analysis of intersection-accidentfrequencies, Journal of Transportation Engineering, Vol. 122, No. 2, 1996, pp. 105-113.

6. Lord, D. The Prediction of Accidents on Digital Networks: Characteristics and Issues Related to the Application of Accident Prediction Models. Ph.D. Dissertation. Department of Civil Engineering, University of Toronto, Toronto, Ontario, 2000.

7. Ivan, J.N., C. Wang, and N. R. Bernardo. Explaining Two-Lane Highway Crash Rates Using Land Use and Hourly Exposure. *Accident Analysis & Prevention*, Vol. 32, No. 6, 2000, pp. 787-795.

8. Lyon, C., J. Oh, B. N. Persaud, S. P. Washington, and J. Bared. Empirical Investigation of the IHSDM Accident Prediction Algorithm for Rural Intersections. *Transportation Research Record 1840*, 2003, pp. 78-86.

9. Miaou, S.-P., and D. Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840*, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 31-40.

10. Oh, J., C. Lyon, S. P. Washington, B.N.Persaud, and J. Bared. Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840*, Transportation Research Board of the National Academies, Washington, D.C., 2003, pp 41-49.

11. Lord, D., A. Manar, and A. Vizioli, Modeling Crash-Flow-Density and Crash-Flow-V/C Ratio for Rural and Urban Freeway Segments. *Accident Analysis & Prevention*, Vol. 37, No. 1, 2005, pp. 185-199.

12. Miaou, S.-P., and J. J. Song. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion and spatial dependence. *Accident Analysis and Prevention*, Vol. 37, No. 4, 2005, pp. 699-720.

13. Tunaru, R. Hierarchical Bayesian Models for Multiple Count Data, Austrian Journal of Statistics, Vol. 31, No. 2&3, 2002, pp.221-229.

14. Bijleveld, F. D. The Covariance Between the Number of Accidents and the Number of Victims in Multivariate Analysis of Accident Related Outcomes. *Accident Analysis and Prevention*, Vol. 37, No. 4, 2005, pp. 591-600.

15. Song, J. J., M. Ghosh, S. Miaou, and B. Mallick. Bayesian Multivariate Spatial Models for Roadway Traffic Crash Mapping. *Journal of Multivariate Analysis*, Vol. 97, 2006, pp. 246-273.

16. Ma, J., and K. M. Kockelman. Bayesian Multivariate Poisson Regression for Models of Injury Count, by Severity. In *Transportation Research Record: Journal of the Transportation Research Board, No.* 1950, Transportation Research Board of the National Academies, Washington, D.C., 2006, pp. 24-34.

17. Tsionas, E. G. Bayesian Multivariate Poisson Regression. *Communications in Statistics—Theory and Methods*, Vol. 30, No. 2, 2001, pp. 243-255.

18. Karlis, D., and L. Meligkotsidou. Multivariate Poisson Regression with Covariance Structure, *Statistics and Computing*, Vol. 15, 2005, pp. 255-265.

19. Chib, S., and R. Winkelmann. Markov Chain Monte Carlo Analysis of Correlated Count Data. *Journal of Business & Economic Statistics,* Vol. 19, 2001, pp. 428-435.

20. Winkelman, R. *Econometric Analysis of Count Data* (4th ed.). New York: Springer, 2003.

21. Tierney, L. Markov Chains for Exploring Posterior Distributions. *Annals of Statistics*, Vol. 22, No. 4, 1994, pp. 1701-1762.

22. Gilks, W. R., S. Richardson, D. J. Spiegelhalter. *Markov chain Monte Carlo in practice.* Chapman & Hall, London, 1996.

23. Liu, J. S. *Monte Carlo Strategies in Scientific Computing.* Springer: New York, 2001.

24. The MathWorks, Inc. MATLAB Neural Network Toolbox 5. Natick, MA, 2006.

25. Anderson, T. W. *An introduction to Multivariate Statistical Analysis* (2nd ed.). New York: Wiley, 1984.

26. Chib, S., and E. Greenberg. Understanding the Metropolis-Hastings Algorithm, *American Statistician*, Vol. 49, No. 4, 1995, pp. 327-335.

27. SAS. Version 9 of the SAS System for Windows. SAS Institute Inc., Cary, NC, 2002.

28. Schluter, P. J., J. J. Deely, and A. J. Nicholson. Ranking and Selecting Motor Vehicle Accident Sites by Using a Hierarchical Bayesian Model, *The Statistician*, Vol. 46, No. 3, 1997, pp. 293-316.

29. Miaou, S.-P., R. P. Bligh, and D. Lord. Developing Median Barrier Installation Guidelines: A Benefit/Cost Analysis using Texas Data. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1904*, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 3-19.

30. Lord, D. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention*, Vol. 38(4), 2006, pp. 751-766.

31. Lord, D., and L. F. Miranda-Moreno. Effects of Low Sample Mean Values and Small Sample Size on the Estimation of the Fixed Dispersion Parameter of Poisson-gamma Models for Modeling Motor Vehicle Crashes: A Bayesian Perspective. Presented at the 86th Annual Meeting of the TRB, Washington, D.C., 2006.