Analytic Methods in Accident Research 12 (2016) 18–34

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

journal homepage: www.elsevier.com/locate/amar

Bayesian nonparametric modeling in transportation safety studies: Applications in univariate and multivariate settings

Shahram Heydari[a,*], Liping Fu[a,b], Lawrence Joseph[c], Luis F. Miranda-Moreno[d]

[a] Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 3G1
[b] Intelligent Transportation Systems Research Center, Wuhan University of Technology, Mailbox 125, No. 1040 Heping Road, Wuhan, Hubei 430063, China
[c] Department of Biostatistics and Epidemiology, McGill University, 687 Pine Avenue West, Montreal, Canada
[d] Department of Civil Engineering and Applied Mechanics, McGill University, 817 Sherbrooke St. W., Montreal, Quebec, Canada H3A 2K6

ARTICLE INFO

Keywords:
Unobserved heterogeneity
Generalized linear models
Dirichlet process mixture models
Latent class models
Multivariate models

ABSTRACT

In transportation safety studies, it is often necessary to account for unobserved heterogeneity and multimodality in data. The commonly used standard or over-dispersed generalized linear models (e.g., negative binomial models) do not fully address unobserved heterogeneity, assuming that crash frequencies follow unimodal exponential families of distributions. This paper employs Bayesian nonparametric Dirichlet process mixture models demonstrating some of their major advantages in transportation safety studies. We examine the performance of the proposed approach using both simulated and real data. We compare the proposed model with other models commonly used in road safety literature including the Poisson-Gamma, random effects, and conventional latent class models. We use pseudo Bayes factors as the goodness-of-fit measure, and also examine the performance of the proposed model in terms of replicating datasets with high proportions of zero crashes. In a multivariate setting, we extend the standard multivariate Poisson-lognormal model to a more flexible Dirichlet process mixture multivariate model. We allow for interdependence between outcomes through a nonparametric random effects density. Finally, we demonstrate how the robustness to parametric distributional assumptions (usually the multivariate normal density) can be examined using a mixture of points model when different (multivariate) outcomes are modeled jointly.

1. Introduction

Generalized linear models (McCullagh and Nelder, 1989; Zeger and Karim, 1991) have been extensively used in analyzing road safety data, conveniently handling crash data through a linear relationship between covariates and log-transformed outcomes such as crash frequencies. Indeed, over-dispersed generalized linear models such as Poisson mixtures (e.g., negative binomial or Poisson-gamma, Poisson-lognormal, etc.) constitute the mainstream approach to account for heterogeneity in crash data (Persaud, 1994; Hauer, 1997; Milton and Mannering, 1998; Karlaftis and Tarko, 1998; Shankar et al., 2003; Ukkusuri et al., 2012). The over-dispersed generalized linear model assumes that crash data follow a unique exponential density. Nevertheless, crash data may arise from a collection of widely differing sub-populations, so that over-dispersed generalized linear models do not fully account for

[*] Corresponding author.
E-mail addresses: shahram.heydari@uwaterloo.ca (S. Heydari), lfu@uwaterloo.ca (L. Fu), lawrence.joseph@mcgill.ca (L. Joseph), luis.miranda-moreno@mcgill.ca (L.F. Miranda-Moreno).

http://dx.doi.org/10.1016/j.amar.2016.09.001
Received 17 January 2016; Received in revised form 8 September 2016; Accepted 8 September 2016
Available online 08 October 2016

unobserved heterogeneity.

As discussed in Mukhopadhyay and Gelfand (1997), compared to over-dispersed generalized linear models, a more comprehensive approach to model heterogeneity would be the finite mixture or latent class models. As Park and Lord (2009) stated, “the mixture model can help provide the nature of the over-dispersion in the data.” Accordingly, a number of road safety studies have recently employed finite mixture models to analyze crash frequency data or differing injury-severity levels (Park and Lord, 2009; Xiong and Mannering, 2013; Zou et al., 2014; Cerwick et al., 2014; Shaheed and Grikitza, 2014).

One important limitation to finite mixture models is that the number of latent components must be prespecified before analyzing the data, but the analyst often does not know the underlying structure of the data a priori. To select the optimal number of components, different models with varying numbers of components must be fit to the data and the one providing the best fit chosen. In practice, a limited number of latent components are usually considered in finite mixture modeling, and the exact number of components may remain uncertain, both of which can compromise the results. In this regard, Behnood et al. (2014) argue that such a limited number of components may result in inadequate approximation of the heterogeneity. For further discussion related to finite mixture modeling, see Mannering et al. (2016).

Another approach in overcoming unobserved heterogeneity in crash data is based on random parameter models such as a random parameter negative binomial model (Anastasopoulos and Mannering, 2009, 2016; Wu et al., 2013; Chen and Tarko, 2014; Mannering and Bhat, 2014; Coruh et al., 2015; Anastasopoulos 2016). In random parameter models, different sets of parameters are estimated for different observations or groups of observations. Therefore, the effect of covariates (contributing factors) is not fixed across all data, but is rather assumed to have a distribution across heterogeneous subsets. While standard random parameter models are limited in their restrictive distributional assumptions, further extensions such as the heterogeneity-in-means approach (Venkataraman et al., 2014) are possible to better address heterogeneity. As discussed in Mannering and Bhat (2014), however, an important limitation to random parameter models is that the analyst must prespecify groupings of observations across which parameters vary. As a consequence, unknown groupings that might exist due to unobserved features are ignored.

Studies in fields such as econometrics have employed finite mixture random parameter models to overcome some of the above issues. This approach relaxes the homogeneity assumption in each latent component of the mixture. In other words, model parameters can vary within each latent component. To our knowledge, such an approach has not been employed in modeling crash frequency data. In road safety literature, Xiong and Mannering (2013) adopted a finite mixture random parameter model to examine the effects of guardian supervision on adolescent driver-injury severities. While such an approach captures unobserved heterogeneity, similar to finite mixture models, the need to prespecifying a limited number of latent components is a shortcoming. For a comprehensive discussion on unobserved heterogeneity in road safety data see Mannering et al. (2016).

Given the above limitations, this paper discusses an alternative, a more flexible Bayesian semiparametric generalized linear model (Escobar and West, 1998; Walker et al., 1999; Neal, 2000; Gelfand and Kottas, 2002; Muller and Quintana, 2004; Hjort et al., 2010). While this approach has been applied in other fields (Mukhopadhyay and Gelfand, 1997; Kleinman and Ibrahim, 1998; Ohlssen et al., 2007; Jara et al., 2007; Muller et al., 2007; Dhavala et al., 2010), applications in transportation research or road safety studies are rare (Heydari et al., 2016; Shirazi et al., 2016; Yu et al., 2016). For example, Heydari et al. (2016) used a Dirichlet process mixture model in a multilevel setting (a form of latent class multilevel model) in which sites were nested within different regions. Bayesian nonparametric models are flexible in the sense that the number of parameters is not fixed and can vary according to data complexity (Gershman and Blei, 2012), taking advantage of Dirichlet process mixtures (Ferguson, 1973; Antoniak, 1974). These models relax restrictive parametric assumptions of conventional modeling approaches and allow identification of latent components (Escobar and West, 1998). Interestingly, the number of latent components can be inferred from the data as part of the analysis.

While most transportation safety researchers have used univariate count models, several road safety researchers have recently employed multivariate models (Ma et al., 2008; Anastasopoulos et al., 2012; Lee et al., 2015; Zhan et al., 2015; Serhiyenko et al., 2016; Barua et al., 2016; Mothafer et al., 2016). These models analyze different injury-severity levels or crash types simultaneously, thereby accounting for correlation, caused by unmeasured or unknown covariates, among outcomes. When such correlation exists, multivariate models provide more accurate estimates and predictions compared to univariate models. For further discussion related to multivariate modeling in transportation safety studies, see Mannering and Bhat (2014) and Mannering et al. (2016). In multivariate settings, an often overlooked issue is the sensitivity to parametric assumptions, with the multivariate normal density almost always defining the dependence structure between outcomes. Dirichlet process mixtures can examine the robustness of this parametric assumption, and can be retained when parametric assumptions do not hold (Muller et al., 1996, 2007; Jara et al., 2007).

The goal of this research is to demonstrate the use of Dirichlet process mixture models in both univariate and multivariate settings. In univariate settings, we show, using both simulated and real data, how the proposed model can examine and relax restrictive parametric assumptions, and eventually capture unobserved heterogeneity. We also compare the adopted model with some of the most commonly used models for count data such as the Poisson-gamma (negative binomial) model, the finite-mixture Poisson-gamma model, and the random intercept model. In multivariate settings, we investigate departures from parametric assumptions and demonstrate how the robustness to standard assumptions can be verified.

We utilize two simulated datasets and three case studies. In defining our models, we follow the methodology discussed in Mukhopadhyay and Gelfand (1997) and Ohlssen et al. (2007) using models that can be estimated in WinBUGS (Lunn et al., 2000). Section 2 describes the methodological framework; Section 3 discusses prior elicitation and model computation; Section 4 discusses model selection and performance measures; Section 5 demonstrates the problem that may arias due to multimodality in model parameters using simulated data; Section 6 exposes the data; Section 7 discusses the results of data analyses; and Section 8 provides conclusions and a summary of the paper.

2. Methodology

In this section we first discuss the realization of a Dirichlet process, followed by a brief overview of the most commonly used generalized linear models in road safety literature. We then describe the proposed model by applying a Dirichlet process mixture to the standard and over-dispersed generalized linear models for count data. Finally, we extend the multivariate Poisson-lognormal model to a more flexible multivariate Dirichlet process mixture model.

2.1. Realization of a Dirichlet process & Dirichlet process mixtures

Let G_O and κ be a continuous baseline distribution (location of the Dirichlet process) and a positive precision (concentration) parameter, respectively. A Dirichlet process can be notated as

$$G \sim Dirichlet(\kappa G_0) \tag{1}$$

A Dirichlet process is a probability measure on the space of all measures (Mukhopadhyay and Gelfand, 1997; Escobar and West, 1998), where for any finite segment S_1,..., S_n of the parameter space, the vector of probabilities ($G(S_1)$,..., $G(S_n)$) follows a Dirichlet distribution with a vector of parameters ($\kappa G_0(S_1)$,..., $\kappa G_0(S_n)$) (Escobar and West, 1998; Muller and Quintana, 2004; Ohlssen et al., 2007). This can be denoted as

$$(G(S_1),\ldots,G(S_n)) \sim Dirichlet(\kappa G_0(S_1),\ldots,\kappa G_0(S_n)) \tag{2}$$

The concentration parameter κ indicates the variability of a Dirichlet process around its baseline distribution. A low value of κ indicates that G can be far from G_0, and vice versa. Therefore, the model with the above structure can be used as a diagnostic tool to verify the robustness of a parametric assumption (Escobar and West, 1998; Ohlssen et al., 2007). A stick-breaking procedure (Ishwaran and James, 2001; Ohlssen et al., 2007) can be implemented to obtain random density functions drawn from a Dirichlet process. The main aim here is to have a set of random probabilities generated sequentially having a sum of one. Such restriction can be guaranteed by the stick-breaking algorithm that breaks a stick with a unit length into an infinite number of partitions. For a detailed discussion see Ishwaran and James (2001) and Muller and Quintana (2004). The stick-breaking procedure, as discussed in Ohlssen et al. (2007), is briefly described as follows: (i) draw a set of random variables θ_1, θ_2,... from G_0; (ii) draw a set of random variables ξ_1, ξ_2,... from a Beta(1, κ); and (iii) allocate probabilities p_1=ξ_1, p_2=$(1-\xi_1)\xi_2$, p_3=$(1-\xi_1)(1-\xi_2)\xi_3$, ... to θ_1, θ_2, θ_3,..., respectively. Note that the probability p and the expectation E for ξ_1, ξ_2,... (Beta distributed random variables) can be obtained from Eq. (3) and Eq. (4).

$$p(\xi_n) = \kappa \xi_n^{\ \kappa-1} \tag{3}$$

$$E(\xi_n) = (1 + \kappa)^{-1} \tag{4}$$

An infinite mixtures of points, the density function *f(.)* corresponding to G, represents realizations of the Dirichlet process (Muller and Quintana, 2004).

$$f(\bullet) = \sum_{n=1}^{\infty} p_n I_{\theta_n},\ \theta_n \sim G_0 \tag{5}$$

In Eq. (5), I_θ is an indicator function corresponding to any θ. Note that *f(.)*, as defined in Eq. (5), is a discrete random probability model. As discussed in Ohlssen et al. (2007), a truncated Dirichlet process (TDP) can be used to approximate a full Dirichlet process with less computational effort, employing standard Markov chain Monte Carlo (MCMC) methods. To do so, it is necessary to limit the maximum number of possible clusters to C (i.e., substitute ∞ with C). Indeed, the truncation occurs at C; and therefore, G depends also on C, i.e., G ~ TDP(κ, G_0, C). In this truncation, it is necessary to restrict the final probability p_c as in Eq. (6). The choice of C could in part be based on the precision parameter κ and is approximately equal to 5κ+2 (Ohlssen et al., 2007).

$$p_C = 1 - \sum_{n=1}^{C} p_n \tag{6}$$

$$f(\bullet) = \sum_{n=1}^{\infty} p_n I_{\theta_n} \approx \sum_{n=1}^{C} p_n I_{\theta_n} \tag{7}$$

The final form of *f*(.) collapses into a finite mixture model that estimates the posterior density of the number of latent clusters in data. We discuss the specification of baseline distribution and priors for model parameters including the precision parameter in the following sections.

2.2. Standard and over-dispersed generalized linear models

A generalized linear model in its simplest form for count data can be described as follows. Let y_i be the observed outcome of interest (e.g., observed crash frequency) for site i. Let $\boldsymbol{X}$ and $\boldsymbol{\beta}$ be the vectors of covariates (i.e., site characteristics) and respective regression coefficients not including the intercept β_0. Then, the model outcome mean λ_i can be related to the covariates using a logarithmic link function $g(.)$,

$$y_i \mid \boldsymbol{X}_i \sim Poisson(\lambda_i) \tag{8}$$

$$log(\lambda_i) = \beta_0 + \boldsymbol{\beta X}_i \tag{9}$$

The above model does not account for over-dispersion and unobserved heterogeneity. Therefore, an extension can be applied to handle over-dispersion. The most common way to overcome heterogeneity is to include an additive error term ε_i,

$$y_i \mid \boldsymbol{X}_i,\, \varepsilon_i \sim Poisson(\lambda_i) \tag{10}$$

$$log(\lambda_i) = \beta_0 + \boldsymbol{\beta X}_i + \varepsilon_i \tag{11}$$

The above model is an over-dispersed generalized linear model. Depending on the distributional assumption for the error term, the above model results in different Poisson mixture settings. Two common Poisson mixtures often used in road safety literature are the Poisson-gamma (negative binomial) model and the Poisson-lognormal model that are respectively obtained by assuming

$$e^{\varepsilon_i} \mid \varphi \sim gamma(\varphi,\, \varphi);\, where\ \varphi \sim gamma(.\ )$$

and

$$\varepsilon_i \mid v_\varepsilon \sim normal(0, \nu_\varepsilon);\, where\ v_\varepsilon^{-1} \sim gamma(.\ )$$

2.3. Extension to the generalized linear Dirichlet process mixture model

To add flexibility to the standard generalized linear model and as a surrogate to the over-dispersed generalized linear model, a Dirichlet process mixture can be adopted to obtain the generalized linear Dirichlet process mixture model, a form of Bayesian semiparametric generalized linear model. We employ a Dirichlet process mixture over the intercept β_0 to tackle heterogeneity to the location of the mean by allowing multimodality as in finite mixture models (Mukhopadhyay and Gelfand, 1997). We retain the linear form for coefficients $\boldsymbol{\beta}$, which in turn retain their usual interpretations.

Recall that the number of latent components in the generalized linear Dirichlet process mixture model can be estimated as part of the analysis, whereas this number must be prespecified (depending on how a priori uncertain it is) in finite mixture models. Inferring the estimated number of latent sub-populations through a systematic mathematical algorithm is more desirable and methodologically sound, assuming the data support such inferences. Let β_{Or} be the intercept for cluster r (1,2,..., C) and G_0 be a normally distributed baseline distribution for β_{Or} with the mean m_0 and the variance v_0. A generic form of the generalized linear Dirichlet process mixture model can be written as follows:

$$y_i \mid \boldsymbol{X}_i,\, \beta_{0r} \sim Poisson(\lambda_i) \tag{12}$$

$$log(\lambda_i) = \beta_{0r} + \boldsymbol{\beta X}_i \tag{13}$$

$$\beta_{0r} \sim Dirichlet(\kappa G_0) \tag{14}$$

$$G_0 \mid m_0,\, v_0 \sim normal(m_0,\, \nu_0) \tag{15}$$

In this model, the precision parameter κ follows a prior distribution $h(.)$. Therefore, its posterior density is estimated as part of the analysis. Similarly, the posterior density of the number of latent components occupied by observations in the data is inferred from data. We discuss the specification of priors in Section 3.

2.4. Extension to the over-dispersed generalized linear Dirichlet process mixture model

After accounting for heterogeneity in data through the generalized linear Dirichlet process mixture model, some extra variability may still exist in some datasets. To account for extra variability, it is possible to use a Dirichlet process mixture over the over-dispersed generalized linear model discussed in Section 2.2. Doing so, besides accounting for over-dispersion by allowing for a flexible model resulting in a mixture of points (in contrast to the parametric unimodal distribution), the remaining variability is accounted for by the error term. As in the generalized linear Dirichlet process mixture model (discussed in Section 2.3), we adopt the method suggested by Mukhopadhyay and Gelfand (1997) in which the authors use a Dirichlet process mixture over the intercept. As discussed previously, this allows maintaining the convenient form of the conventional over-dispersed generalized linear models for the covariates. Given the above notation, the over-dispersed generalized linear Dirichlet process mixture model can be specified as

$$y_i \mid \boldsymbol{X}_i,\, \beta_{0r},\, \varepsilon_i \sim Poisson(\lambda_i) \tag{16}$$

$$log(\lambda_i) = \eta_r + \boldsymbol{\beta X}_i + \varepsilon_i \tag{17}$$

$$\beta_{0r} \sim Dirichlet(\kappa G_0) \tag{18}$$

$$G_0 \mid m_0, v_0 \sim normal(m_0, \nu_0) \tag{19}$$

To circumvent identifiability issues the mean of the error term ε_i is fixed to be equal to zero; i.e., ε_i ~ normal(0, υ_ε).

2.5. Dirichlet process mixing in multivariate settings

Given the above notation, a conventional multivariate Poisson-lognormal model to jointly analyze j correlated outcomes can be defined as

$$y_{ij} \mid \boldsymbol{X}_{ij}, \varepsilon_{ij} \sim Poisson(\lambda_{ij}) \tag{20}$$

$$log(\lambda_{ij}) = \beta_{0j} + \boldsymbol{\beta}_j \boldsymbol{X}_{ij} + \varepsilon_{ij} \tag{21}$$

$$\varepsilon_{ij} \sim MVN(0, \Sigma) \tag{22}$$

The vector of error terms ε, which accounts for correlation among outcomes, is generated from a multivariate normal (MVN) distribution with covariance matrix Σ, where the inverse of Σ is assumed to follow a Wishart distribution. For details on the structure of the multivariate model see, for example, Ma et al. (2008) and El-Basyouny and Sayed (2009).

To extend the standard multivariate model to the Dirichlet process mixture multivariate model, the error term, ε_{ij}, can be included in the intercept term to allow variation across observations with respect to the intercept. Let's m_j denote the mean of the varying intercept. We can thus write

$$log(\lambda_{ij}) = \beta_{0ij} + \boldsymbol{\beta}_j \boldsymbol{X}_{ij} \tag{23}$$

$$\beta_{0ij} \sim MVN(m_j, \Sigma) \tag{24}$$

The Bayesian nonparametric allows relaxing of the parametric assumption for the jointly distributed error terms (here correlated random intercepts). The Dirichlet process multivariate model uses a parametric density that is usually a multivariate normal distribution as its baseline density G_0 and then allows departures from this parametric assumption. Note that the same analogy can be used in simultaneous equation modeling or endogenous settings to relax restrictive parametric assumptions. The model can thus be defined as

$$y_{ij} \mid \boldsymbol{X}_{ij}, \varepsilon_{ij} \sim Poisson(\lambda_{ij}) \tag{25}$$

$$log(\lambda_{ij}) = \beta_{0rj} + \boldsymbol{\beta}_j \boldsymbol{X}_{ij} \tag{26}$$

$$\beta_{0rj} \sim Dirichlet(\kappa G_0) \tag{27}$$

$$G_0 \sim MVN(m_{0j}, \Sigma) \tag{28}$$

where m_{0j} is the mean of the outcome j for the multivariate normal baseline G_0. The correlated parameters (random intercepts) are modeled as a mixture of points. While one can allocate a Dirichlet prior on the error term without involving the intercept, this results in further complexity as the mean of the Dirichlet cannot be equal to 0.

3. Prior specification and model computation

Non-informative priors were set for regression coefficients and the mean of the baseline distribution. In particular, we used normally distributed priors with mean zero and a large variance. We used gamma (0.01,0.01) priors for the inverse variance. See Heydari et al. (2014) for a detailed discussion related to prior specification in road safety studies in univariate settings. The parameterization of the Wishart density in the multivariate setting was selected to reflect a vague prior specification (El-Basyouny and Sayed, 2009). With respect to the baseline distribution, note that we did not fixed the baseline parameters (mean and variance) in this study. Instead, we used vague hyper priors on these parameters and let the model estimate them. It is important to consider that if a baseline does not support the range of a dataset, the model would not be able to make proper posterior inference. Using vague hyper priors for the baseline helps prevent such condition.

One also needs to select a prior for the Dirichlet precision parameter κ, an important choice, since its posterior density is critical in deciding how closely a parametric distributional assumption holds. A gamma or uniform prior is usually selected for this prior. For example, Ohlssen et al. (2007) chose a uniform prior with lower and upper limits of 0.3 and 10, respectively, while Ishwaran and James (2001) suggested a gamma prior with shape and scale parameters equal to 2. For a detailed discussion in this regard, see Ishwaran (2000), Ohlssen et al. (2007), Dorazio (2009), and Murugiah and Sweeting (2012).

The prior for k is related to the maximum number of possible clusters C, discussed in Section 2.1. For the data analyzed here, we considered C to be equal to 52 (based on 5α+2, where the upper limit of α is set to 10), a relatively large number, so that we were able to approximate an infinite mixture of points. Based on the maximum number of components (i.e., 50) and discussion provided in Section 2.1, we initially used a uniform prior for κ with a lower limit of 0.3 and an upper limit of 10 for the vehicle injury and the highway 401 datasets. Such values allow small and large values of κ while avoiding problems relating to the calculation of p_n (defined in Section 2.1). We also verified the sensitivity to the prior choice for the Dirichlet precision parameter by choosing an upper limit of 20 and 100 for the vehicle injury dataset and the highway 401 dataset, respectively. The results in this regard are reported in Section 7.

For the grade crossing dataset having a high proportion of zero crashes, we first used a uniform prior with lower and upper limits of 0.3 and 10, respectively. However, the dataset, being limited, couldn't provide much information about the Dirichlet precision parameter. The estimated interval around this parameter varied from 0.7 to 8.4 that is quite a large interval, which is similar in range to the specified prior. We then used a gamma prior with the shape and scale parameters set to one, a somewhat more informative prior. While the selected gamma prior is inclined to result in small values of the precision parameter, it has a relatively heavy tail that also allows larger values although the probability for such values is small. Note that almost 90% of the crossings in this dataset did not experience any crash over a 6-year period, so lower values of the precision parameter were more plausible, justifying that prior choice. In the grade crossings dataset, the density of observed crashes bunches up at zero with around 90% of observations concentrated at this peak. It is doubtful that the random intercepts follow a normal distribution, but we rather expect these to concentrate near zero, and with a limited number of latent components. The gamma(1, 1) supported these expectations and resulted in a better and quicker mixing of chains in the MCMC algorithm.

WinBUGS (Lunn et al., 2000) was used to generate MCMC samples for Bayesian posterior inference. For the crossings dataset, we used two chains with 100,000 iterations each. The first 20,000 iterations were considered as burn-in and convergence requirements. Posterior estimates were thus obtained using 80,000 iterations or 160,000 samples. For the Vehicle-injury dataset, two chains with 80,000 iterations were considered among which the first 20,000 were discarded for burn-in and model convergence, so 120,000 samples were utilized for inference. This was sufficient for low Monte Carlo errors. The same number were used in the multivariate setting and for the highway 401 dataset. History plots, trace plots, and the Gelman-Rubin statistic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) were used to ensure that convergence was reached.

4. Model selection and performance measures

As discussed in Section 2.1, Dirichlet process mixture models can be used to check how closely a parametric assumption might hold. If a parametric assumption (for example, a normality assumption for random intercepts) seems far from true, there is justification to avoid using that parametric model. At this point, there is no further need for other model selection methods for that model, which has been ruled out. Nevertheless, we do discuss some model fitting criteria here as an extra piece of information; for example, to show that how the predictive capability of a model can be affected by an assumption that does not hold.

The deviance information criterion (DIC) is usually used for model selection in Bayesian crash data analysis. However, the DIC is sensitive to different parameterizations (Geedipally et al., 2014) and of questionable use in case of multimodal posteriors (Washington et al., 2010). A discussion about some of the limitations associated with the DIC can be found in Carlin and Louis (2008). In this article, we used cross-validation predictive densities (Gelfand, 1996; Mukhopadhyay and Gelfand, 1997; Vehtari and Lampinen, 2002; Ntzoufras, 2009) to compute conditional predictive ordinates (CPOs) that in turn allow estimating the log pseudo marginal likelihood (LPML) and the pseudo Bayes factor (PBF). The cross-validation method compares alternative models in terms of their predictive abilities. A relatively detailed discussion in this regard is provided in Ntzoufras (2009), Carlin and Louis (2008), and Heydari et al. (2016). Here, we briefly discuss the main components of this method.

Suppose Y_i is the ith observation, T stands for the total number of iterations in an MCMC simulation, and ψ represents the estimated model parameters. For each observation, the CPO can be estimated as

$$CPO_i = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{f(Y_i \mid \psi^{(t)})} \right)^{-1} \tag{29}$$

The product of CPOs across all subjects gives the pseudo marginal likelihood (PML), from which the PBF of comparing model 1 against model 2 can be obtained from

$$PBF = PML_{model1}/PML_{model2} \tag{30}$$

Alternatively, the LPML (Gelfand et al., 1992), given in Eq. (31), is easier to calculate.

$$LPML = log\ \{ \prod_{i=1}^{l} CPO_i \} = \sum_{i=1}^{l} log\ (CPO_i) \tag{31}$$

The model with the highest LPML indicates a better fit to the data. We also examined the performance of the proposed model in terms of its ability to replicate a high proportion of zero crashes as in the grade crossing dataset. This posterior predictive check is based on a selected statistic of interest as discussed in Rubin (1984). To implement, we first replicated crash observations based on estimated expected crash frequencies inside the MCMC algorithm. A Bayesian p-value (Gelman et al., 1996) then compares the proportions of zeros in replicated and observed data. A p-value of 0.5 indicates a perfect similarity between the observed and replicated data.

5. A simulated example

One important advantage of simulated data is that the true parameters and the structure of the data are known, so that one can evaluate the accuracy of posterior inferences from any model. In this paper, two simulated data were used: (1) a dataset with bimodal intercepts concentrated at two distinct values; and (2) a dataset with intercepts concentrated at a single value. For the first data simulation scenario, we generated two crash datasets, both with 100 observations, and varying only in their intercepts. The total number of generated observations is 200, which is sufficient here since the simulated data are only intended as an example to illustrate how Dirichlet process mixture models work. If we were instead aiming to provide detailed properties of a new model via a simulation study, then a larger sample might have been indicated. Data were generated from a Poisson distribution with expected crash frequency, a function of a single hypothetical covariate, say, traffic exposure. In particular, we generated data as follows:

$$y_i \mid x_i \sim Poisson(\lambda_i) \tag{32}$$

$$log(\lambda_i) = \beta_0 + \beta x_i \tag{33}$$

To create the above scenario, we randomly selected 100 observations from a railway grade crossing dataset, described in Section 6, where traffic exposure, x, was known. Since we assumed that covariates and their effect are identical, we used the same set of observations selected above to build the second subset containing 100 observations. We set the value of β (in Eq. (33)) to be 0.492. To generate crashes based on the model structure defined above, we set the intercept to be equal to −4 for the first subset and 3 for the second subset. Both subsets were then combined to create a single dataset with 200 observations. Doing so, both subsets were identical except in their two distinct intercepts. We then analyzed the simulated data using the proposed Dirichlet process mixture model, the finite-mixture Poisson-gamma model with two and three latent components, the standard Poisson-gamma (negative binomial) model, and the random intercepts (random effects) Poisson model. Readers are referred to Zou et al. (2014) and Chen and Tarko (2014) for modeling details relating to finite mixture and random effects models, respectively.

The results are reported in Table 1. The Dirichlet process mixture model correctly identifies the two clusters in the simulated data (see Fig. 1(a)). It also accurately estimates other model parameters. The conventional finite mixture model with two components also performs well. The intercept and beta coefficient are estimated accurately, and over-dispersion parameters for each component are estimated to be very large indicating that the distribution of crash frequency in each subset is Poisson. Note that a large value of over-dispersion is expected here since we generated each subset from a simple Poisson distribution. The finite mixture model with three

Table 1
Summary of the posterior densities for the simulated data.

	Posterior Mean	**Std. Dev.**	**95% Credible intervals**	
			2.50%	**97.50%**
Dirichlet process mixture Poisson model				
Intercept mean	−0.601	0.270	−1.143	−0.085
Intercept variance	13.160	1.630	12.020	16.820
Covariate coefficient	0.491	0.001	0.489	0.494
Baseline mean	−1.423	3.930	−9.457	6.559
Baseline Std. Dev.	5.999	2.107	2.428	9.742
Precision parameter k	0.706	0.382	0.312	1.698
Finite-mixture Poisson-gamma model with 2 components				
Intercept (component 1)	−4.148	0.452	−5.049	−3.283
Intercept (component 2)	2.998	0.018	2.963	3.035
Covariate coefficient (component 1)	0.492	0.045	0.404	0.580
Covariate coefficient (component 2)	0.492	0.002	0.488	0.496
Over-dispersion (component 1)	47.330	55.510	4.844	207.900
Over-dispersion (component 2)	794.600	207.900	460.100	1269.000
Finite-mixture Poisson-gamma model with 3 components				
Intercept (component 1)	−2.557	0.615	−4.202	−1.868
Intercept (component 2)	2.998	0.019	2.962	3.035
Intercept (component 3)	1.953	1.472	−0.850	4.999
Covariate coefficient (component 1)	0.368	0.063	0.279	0.495
Covariate coefficient (component 2)	0.492	0.002	0.488	0.497
Covariate coefficient (component 3)	0.527	0.744	−0.126	2.605
Over-dispersion (component 1)	18.160	41.040	0.046	134.700
Over-dispersion (component 2)	798.600	210.000	459.100	1275.000
Over-dispersion (component 3)	0.022	0.009	0.012	0.041
Standard Poisson-gamma model				
Intercept	2.758	0.518	1.841	3.842
Covariate coefficient	0.442	0.057	0.328	0.545
Over-dispersion	0.119	0.011	0.098	0.142
Random intercept Poisson model				
Intercept mean	−4.413	0.469	−5.293	−3.442
Intercept variance	22.030	2.891	17.040	28.340
Covariate coefficient	1.010	0.043	0.902	1.071

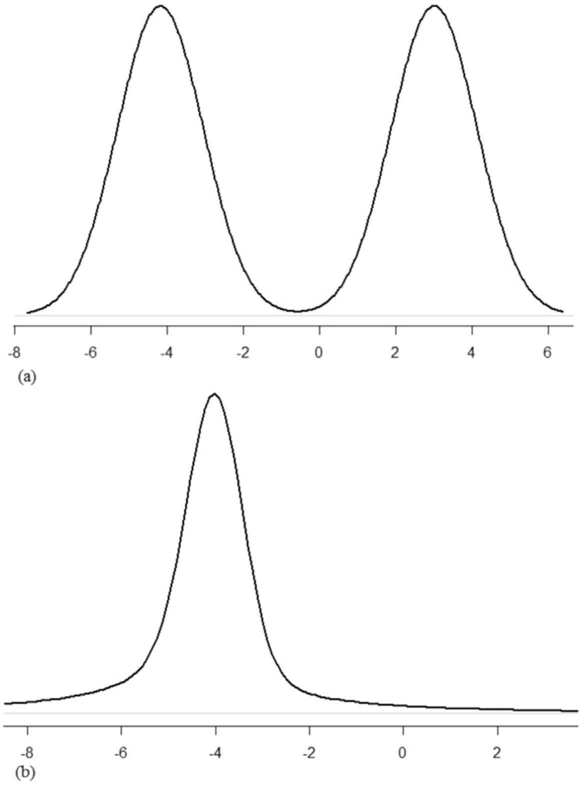

Fig. 1. Kernel density plot of the posterior density for the intercept for two simulated datasets: (a) scenario 1; and (b) scenario 2.

components (wrong number of components) works less well, with biased estimates, and similarly biased results are obtained from the Poisson-gamma and the random intercepts models. The Poisson-gamma model assumes that the intercept is fixed, while the random intercepts model allows the intercept to vary, but following a normal density. With neither assumption holding, it is not surprising that these models do not work well. Conversely, the Dirichlet process mixture model works well when these assumptions do not hold.

For the second data simulation scenario, we randomly selected 200 observations from a grade crossing dataset. Similar to the first scenario, this simulated dataset was generated using (32) and (33) with only one covariate; i.e., traffic exposure. Model parameters β and β_0 were set to be 0.492 and −4, respectively. The data were generated from a simple Poisson distribution with fixed parameters, so that there is no multimodality in any component of the data. We first analyzed this simulated dataset using the standard negative binomial model. This model estimated β and β_0 to be 0.488 and −4.06, respectively. The over-dispersion parameter was estimated to be 65.73 indicating that the distribution of the data is close to the simple Poisson. We then analyzed this dataset using the Dirichlet process model; the results were found to be very similar to those obtained from the negative binomial model. In particular, the Dirichlet process model estimated β and β_0 to be 0.486 and −4.066, respectively. A kernel density plot of the posterior density for the intercept, obtained from the Dirichlet process model, is illustrated in Fig. 1(b) showing a unimodal density concentrated at −4, as expected. Similar to the first scenario, the Dirichlet process model accurately estimated the model parameters and the structure of the data (here, the form of the intercept). In both scenarios the Dirichlet process model performed well. This is a valuable property of the Dirichlet process mixture models that can adjust themselves to the complexity of any data (Gershman and Blei, 2012).

6. Data

Two datasets were used to illustrate the generalized linear Dirichlet process mixture model and the over-dispersed generalized linear Dirichlet process mixture model in univariate settings. The first dataset contains vehicle-injury counts for 647 signalized intersections in Montreal from 2003 to 2008. This dataset is highly over-dispersed and characterized by a relatively large mean value. The vehicle-injury data were provided by ambulance services. Other information such as geometric characteristics (number of lanes, presence of median, etc.), built environment characteristics (population, land use, presence of bus and subway stations, etc.), and traffic control characteristics (signal type, etc.) were obtained from various sources. Summary statistics of this dataset are reported in Table 2. Note that the vehicle-injury dataset has an average mean value of 4.6 injuries in a six-year period. Among 647 signalized intersections, 143 (22.10%) were three-leg intersections, 458 (70.79%) were in the proximity of bus stops, and 364 (56.26%) were in a distance of less than 400 m from a school. The number of intersections with at least one raised median was 290

Table 2
Summary statistics for the vehicle-injury data (647 observations).

Variable	Mean	Std. Dev.	Min	Max
Through AADT	19,467.96	11,084.39	1,790.00	76,525.00
Left-turning AADT	2,602.72	2,641.86	0	23,843.00
Right-turning AADT	2,668.01	2,697.45	0	23,792.00
Ratio of pedestrians & bikes over total AADT	0.226	0.467	0.003	7.574
Total number of lanes for all approaches	6.90	2.60	2.00	16.00
Number of subway stations in 400 m	0.44	0.70	0.00	4.00
Three-leg (1 if three-leg intersection; 0 otherwise)	0.22	0.42	0.00	1.00
Bus stop (1 if present in 50 m; 0 otherwise)	0.71	0.46	0.00	1.00
Raised median (1 if present; 0 otherwise)	0.47	0.50	0	1.00
School (1 if present in 400 m; 0 otherwise)	0.56	0.50	0.00	1.00
Vehicle-injury frequency	4.60	6.37	0.00	58.00

Table 3
Summary statistics for the grade crossing data (6617 observations).

Variable	Mean	Std. Dev.	Min	Max
Train flow (number of trains daily)	11.071	12.976	0.100	162.000
Vehicle flow (AADT)	3,082.396	5,636.744	1.000	71,500.000
Exposure (product of train flow and vehicle flow)	29,695.710	94,428.280	0.270	300,0000.000
Train ratio (ratio of train flow to vehicle flow)	0.170	1.854	0.000	54.000
Number of rail tracks	1.292	0.612	1.000	7.000
Number of lanes	2.164	0.671	1.000	7.000
Road speed (speed limit in km/h)	62.333	17.879	5.000	110.000
Train speed (maximum train speed in km/h)	63.910	36.446	1.608	160.800
ln(road speed)*ln(ratio of train flow to vehicle flow)	−20.323	9.471	−56.350	17.314
Track angle (deviation from 90 degrees)	19.496	19.709	0.000	87.000
Gate (1 if gate is present; 0 otherwise)	0.364	0.481	0.000	1.000
Whistle prohibition (1 if prohibited; 0 otherwise)	0.130	0.336	0.000	1.000
Urban (1 if located in urban area; 0 otherwise)	0.354	0.478	0.000	1.000
Ont./Qc. (1 if located in Ontario or Quebec)	0.578	0.494	0.000	1.000
Pac./Atl. (1 if located in Pacific or Atlantic region)	0.154	0.361	0.000	1.000
Crash frequency	0.080	0.317	0.000	4.000

Table 4
Summary statistics for the highway 401 data (418 observations).

Variable	Mean	Std. Dev.	Min	Max
AADT all vehicles	80,369.420	95,760.440	14,499.940	442,900.300
AADT commercial vehicles	14,383.640	6,890.880	4,864.000	42,075.500
Percentage of commercial vehicles	29.027	12.300	3.100	49.100
Segment length (km)	1.952	2.061	0.206	12.703
Number of lanes	5.445	2.428	4.000	12.000
Median (inside) shoulder width (m)	1.598	1.194	0.000	5.190
Median width (m)	11.106	6.147	0.600	30.500
Outside shoulder width (m)	3.135	0.285	2.600	4.000
Lane width (m)	3.707	0.301	1.830	5.625
Average horizontal curve degree curvature per km	0.945	1.864	0	16.592
Paved outside shoulder (1 if paved; 0 otherwise)	0.586	0.493	0.000	1.000
Surface type (1 if HCB[a]; 0 otherwise)	0.526	0.500	0.000	1.000
Narrow median shoulder (1 if < 1.8 m; 0 otherwise)	0.629	0.493	0.000	1.000
Property-damage-only crash frequency	18.715	38.257	0.000	336.000
Injury-fatal crash frequency	4.530	9.334	0.000	96.000

[a] HCB stands for high class bituminous pavement.

(44.82%). For further discussion relating to this dataset, see Strauss et al. (2014).

The second dataset is characterized by a very low mean value and excess zero counts. This dataset records crash frequencies at 6617 automated railway grade crossings in Canada. Automated crossings are equipped with flashing lights, bells and/or gates to inform road users about approaching trains. The data were provided by Transportation Safety Board of Canada covering a six-year period from 2008 to 2013 (Heydari and Fu, 2015). A host of independent variables (including geometric and operational attributes) were available in the database, the most important shown in Table 3. We also created three dummy variables to reflect spatial effects to some extent based on similarities observed in an exploratory data analysis phase. The prairie region, consisting of the provinces of

Table 5
Posterior distribution summaries for the vehicle-injury dataset.

	Posterior Mean	**Std. Dev.**	**95% Credible intervals**	
			2.50%	**97.50%**
Over-dispersed Dirichlet process mixture Poisson model				
Intercept mean	−8.746	1.189	−11.290	−6.743
Intercept variance	14.370	15.830	2.796	58.410
ln(through AADT)	0.484	0.099	0.293	0.738
ln(right-turning AADT)	0.240	0.050	0.149	0.354
ln(left-turning AADT)	0.177	0.041	0.091	0.252
ln(ratio of pedestrians & bikes over total AADT)	−0.112	0.040	−0.112	0.040
Presence of bus stop	0.298	0.131	0.048	0.561
Presence of subway station	0.199	0.118	0.001	0.423
Dirichlet baseline mean	−11.090	3.441	−19.010	−5.224
Dirichlet Baseline Std. Dev.	4.984	2.268	1.673	9.552
Dirichlet precision parameter k	1.771	1.372	0.384	5.628
Variance υ_ε (for extra variation)	0.487	0.141	0.125	0.705
Log pseudo marginal likelihood	−1462.670	–	–	–
Finite-mixture Poisson-gamma model with 2 components				
Component 1				
Intercept	−8.205	0.627	−9.593	−7.075
ln(through AADT)	0.645	0.044	0.556	0.722
ln(right-turning AADT)	0.433	0.095	0.268	0.626
ln(left-turning AADT)	−0.129	0.053	−0.248	−0.060
ln(ratio of pedestrians & bikes over total AADT)	0.338	0.092	0.178	0.529
Presence of bus stop	1.782	0.356	1.186	2.517
Presence of subway station	−0.153	0.229	−0.613	0.299
Over-dispersion	0.583	0.095	0.003	0.415
Component 2				
Intercept	−0.153	0.229	−0.613	0.299
ln(through AADT)	0.209	0.149	0.020	0.409
ln(right-turning AADT)	0.095	0.038	0.033	0.171
ln(left-turning AADT)	0.401	0.056	0.293	0.515
ln(ratio of pedestrians & bikes over total AADT)	−0.165	0.045	−0.269	−0.087
Presence of bus stop	−0.120	0.131	−0.395	0.119
Presence of subway station	0.124	0.146	−0.166	0.430
Over-dispersion	4.230	1.277	2.542	7.300
Log pseudo marginal likelihood	−1549.690	–	–	–
Standard Poisson-gamma model				
Intercept	−6.983	0.921	−8.792	−5.180
ln(through AADT)	0.486	0.092	0.306	0.668
ln(right-turning AADT)	0.201	0.035	0.132	0.271
ln(left-turning AADT)	0.189	0.034	0.121	0.256
ln(ratio of pedestrians & bikes over total AADT)	−0.092	0.040	−0.172	−0.014
Presence of bus stop	0.462	0.123	0.218	0.703
Presence of subway station	0.348	0.126	0.101	0.597
Over-dispersion	0.763	0.060	0.654	0.886
Log pseudo marginal likelihood	−1496.86	–	–	–
Random intercept over-dispersed Poisson model				
Intercept mean	−7.649	0.768	−9.066	−6.078
Intercept variance	0.886	0.489	0.047	1.540
ln(through AADT)	0.499	0.083	0.331	0.656
ln(right-turning AADT)	0.214	0.044	0.128	0.302
ln(left-turning AADT)	0.174	0.044	0.089	0.261
ln(ratio of pedestrians & bikes over total AADT)	−0.013	0.048	−0.107	0.080
Presence of bus stop	0.718	0.135	0.455	0.985
Presence of subway station	0.329	0.137	0.059	0.600
Variance υ_ε (for extra variation)	0.490	0.480	0.010	1.408
Log pseudo marginal likelihood	−1512.18	–	–	–

Manitoba, Saskatchewan, and Alberta, was selected as the reference group. Ontario and Quebec formed another group. Finally, the Pacific region (British Columbia) and the Atlantic region (New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island) formed the Pacific/Atlantic region. The crossing dataset had a very low mean crash frequency with almost 90% of crossings experiencing no crash over the aforementioned period. Summary statistics of the crossing data are reported in Table 3.

The third dataset, provided by the Ontario Ministry of Transportation, was used to illustrate the application of Dirichlet process

Table 6
Posterior distribution summaries for the grade crossing crash dataset.

	Posterior Mean	Std. Dev.	95% Credible intervals	
			2.50%	97.50%
Dirichlet process mixture Poisson model				
Intercept mean	−7.693	1.248	−11.250	−6.223
Intercept variance	5.383	14.200	0.139	41.440
ln(traffic exposure)	0.488	0.035	0.420	0.557
ln(train speed)	0.226	0.098	0.036	0.423
ln(road speed)*ln(train ratio)	0.014	0.008	0.001	0.029
Presence of gate	−0.686	0.126	−0.934	−0.437
Ontario/Quebec[1]	−0.913	0.105	−1.120	−0.705
Pacific/Atlantic region[1]	−0.575	0.145	−0.860	−0.292
Dirichlet baseline mean	−8.045	2.785	−15.600	−4.393
Dirichlet Baseline Std. Dev.	3.374	2.286	0.847	9.168
Dirichlet precision parameter k	0.922	0.509	0.323	2.232
Log pseudo marginal likelihood	−1687.65	–	–	–
Standard Poisson-gamma model				
Intercept	−6.631	0.523	−7.651	−5.525
ln(traffic exposure)	0.490	0.037	0.420	0.566
ln(train speed)	0.181	0.101	0.012	0.386
ln(road speed)*ln(train ratio)	0.016	0.008	0.001	0.032
Presence of gate	−0.676	0.126	−0.925	−0.434
Ontario/Quebec	−0.911	0.105	−1.117	−0.711
Pacific/Atlantic region	−0.584	0.147	−0.867	−0.289
Over-dispersion	0.912	0.225	0.599	1.467
Log pseudo marginal likelihood	−1731.720	–	–	–
Random intercept over-dispersed Poisson model				
Intercept mean	−7.364	0.545	−8.595	−6.577
Intercept variance	0.871	0.161	0.586	1.218
ln(traffic exposure)	0.491	0.036	0.422	0.562
ln(train speed)	−0.688	0.128	−0.940	−0.439
ln(road speed)*ln(train ratio)	0.237	0.102	0.062	0.456
Presence of gate	0.013	0.008	0.001	0.029
Ontario/Quebec	−0.914	0.105	−1.121	−0.707
Pacific/Atlantic region	−0.572	0.147	−0.866	−0.290
Variance υ_{ε} (for extra variation)	0.871	0.161	0.586	1.218
Log pseudo marginal likelihood	−1710.760	–	–	–

[1] The Prairie region is the reference region.

mixtures in multivariate settings. This dataset consists of crash data from 418 highway segments in Ontario (highway 401) collected over a 3-year period 2006–2008). Highway 401 connects eastern Ontario (the Quebec boarder) to south west Ontario (the Michigan boarder). This highway is a major roadway in Ontario with a very large number of vehicles passing through it on a daily basis. The crash data are divided into three categories of severities: fatal, injury, and property damage only crashes. Due to limited number of fatal crashes, we divided the crash data into two categories of injury-fatal and property damage only crashes. The dataset does not distinguish between various levels of injury such as incapacitating injury, etc. In addition to the crash data, major roadway-segment information was available. Descriptive statistics of the highway 401 dataset are provided in Table 4. We noticed a higher rate of crashes among segments with a median shoulder width of smaller than 1.80 m during an exploratory data analysis phase. Based on the median (inside) shoulder width, we created a dummy independent variable, here named narrow median shoulder. No information relating to the vertical alignment of segments was available, but we were able to obtain the average horizontal curve degree per kilometer of highway segment.

7. Results and discussion

Posterior inferences are summarized in Table 5, Table 6, and Table 7 . Several independent variables were discarded due to co-linearity issues. In the presence of dummy variables such as the presence of a gate, endogeneity (Kim and Washington, 2006) may exist in the models. While we recognize the general importance of addressing endogeneity in crash modeling, the topic is beyond the scope of this research that focuses more on methodological rather than on empirical aspects. Once we introduce Dirichlet process mixtures in a standard framework, further extensions such as accommodating endogeneity can then follow, the topic deserving a paper on its own.

Table 7
Posterior distribution summaries for the highway 401 dataset.

	Posterior Mean	Std. Dev.	95% Credible intervals	
			2.50%	97.50%
Dirichlet process mixture multivariate Poisson-lognormal model				
Property damage only crashes				
Intercept mean	−10.950	0.291	−11.500	−10.460
Intercept variance	0.703	0.127	0.513	1.006
ln(AADT)	1.267	0.026	1.223	1.316
ln(length)	0.754	0.028	0.701	0.810
Average horizontal curve degree curvature per km	−0.146	0.016	−0.176	−0.113
Narrow median (inside) shoulder	0.160	0.042	0.080	0.249
Multivariate normal baseline mean	−10.980	0.390	−11.760	−10.270
Injury-fatal crashes				
Intercept mean	−12.050	0.415	−12.800	−11.21
Intercept variance	0.322	0.057	0.235	0.457
ln(AADT)	1.291	0.034	1.220	1.354
ln(length)	0.803	0.033	0.739	0.869
Average horizontal curve degree curvature per km	−0.072	0.019	−0.108	−0.034
Median (inside) shoulder width	−0.079	0.019	−0.116	−0.042
ln(median width)	−0.077	0.034	−0.145	−0.011
Paved outside shoulder	−0.245	0.068	−0.377	−0.113
Multivariate normal baseline mean	−12.090	0.448	−12.910	−11.200
Dirichlet precision parameter k	8.752	1.035	6.184	9.961
Correlation between outcomes	0.943	0.030	0.866	0.983
Log pseudo marginal likelihood	−2022.710	–	–	–
Standard multivariate Poisson-lognormal model				
Property damage only crashes				
Intercept mean	−10.720	0.403	−11.290	−9.870
ln(AADT)	1.247	0.036	1.172	1.298
ln(length)	0.748	0.049	0.650	0.842
Average horizontal curve degree curvature per km	−0.146	0.024	−0.193	−0.100
Narrow median shoulder	0.157	0.073	0.014	0.298
Injury-fatal crashes				
Intercept mean	−11.280	0.506	−12.420	−10.400
ln(AADT)	1.232	0.043	1.158	1.327
ln(length)	0.793	0.045	0.706	0.881
Average horizontal curve degree curvature per km	−0.073	0.023	−0.117	−0.028
Median (inside) shoulder width	−0.084	0.024	−0.131	−0.037
ln(median width)	−0.132	0.039	−0.209	−0.054
Paved outside shoulder	−0.238	0.072	−0.380	−0.097
Correlation between outcomes	0.876	0.022	0.829	0.914
Log pseudo marginal likelihood	−2021.390	–	–	–

7.1. The univariate setting

We first analyzed the vehicle-injury data using the standard Poisson-gamma model. We then allowed the intercept to vary across observations (random intercept or random effects model) following a normal distribution. We also analyzed these data using a finite-mixture Poisson-gamma model with two and three components, but since the 3-component model did not improve the fit, only the results for the 2-component model are reported.

We compared these standard models to the Dirichlet process mixture on intercept. The results from the Dirichlet process mixture model showed that the Dirichlet precision parameter k is concentrated at some point close to the lower limit of 0 (Fig. 2). Recall that a low value of κ indicates that G is far from G_O, as discussed in Section 2. Therefore, the normal assumption for intercept is unlikely to hold. That is the 647 random intercepts are not normally distributed, with evidence of multimodality that can be captured in the form of latent clusters. In fact, the Dirichlet process model estimates the posterior median of the number of clusters to be 8 (3, 25). A histogram of the posterior number of clusters is shown in Fig. 3.

The log pseudo marginal likelihoods suggest that the random intercept model does not provide a better fit compared to the Poisson-gamma model. This is similar to the case discussed in Ohlssen et al. (2007). To verify the sensitivity to the initial prior choice for k (a uniform distribution with lower and upper bounds of 0.3, 10, respectively), we analyzed the data using a different prior, k ~ uniform(0.3, 20), and obtained similar results. We also verified the sensitivity to hyper prior choice for the baseline density although this was not a priori of a major concern as the initial hyper priors were selected to be vague. For example, when we changed the variance of the specified hyper prior from 100 to 400, only a minor difference was observed, with the point estimate of k remaining stable changing from 1.771 to 1.784, without any particular change in the form of the posterior.

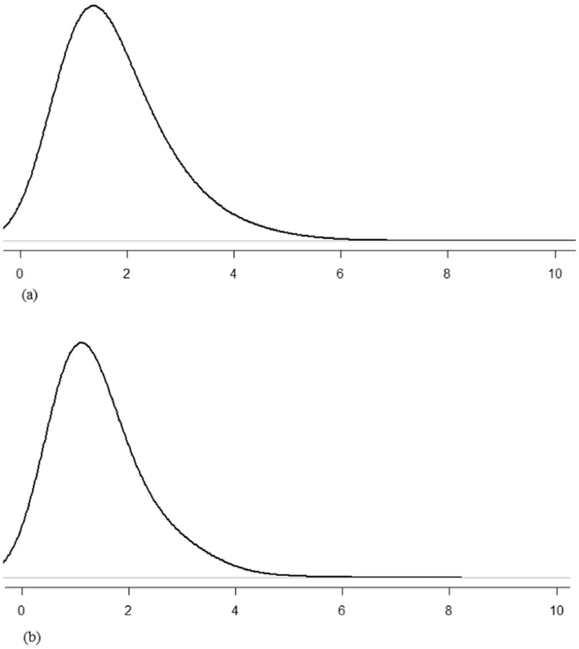

Fig. 2. Kernel posterior density plots of Dirichlet precision parameter k for the vehicle-injury dataset for two different prior densities: (a) k ~ uniform (0.3, 10); and (b) k ~ uniform (0.3, 20).

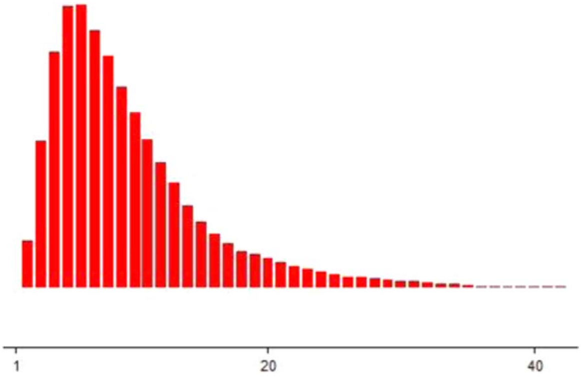

Fig. 3. Histogram of the posterior number of latent clusters for the vehicle-injury dataset.

The vehicle-injury dataset is highly over-dispersed with a relatively high mean value, so that it is not surprising to find models accommodating these features supported by the criteria reported in Table 5. It can be implied from Table 5 that log pseudo marginal likelihoods significantly differ from one model to another. The over-dispersed generalized linear Dirichlet process mixture model provides the highest log pseudo marginal likelihood of −1462.67 followed by the standard Poisson-gamma model, the random intercept model, and then the finite mixture Poisson-gamma model. In comparison to the Poisson-gamma model, for example, we obtain a log pseudo Bayes factor of 34.01 (1496.86–1462.67) that provides support for the over-dispersed Dirichlet process mixture model. Based on the results, through AADT, left-turning AADT, right-turning AADT, the presence of bus stop, and the presence of subway station are positively associated with vehicle-injury counts. However, the ratio of the number of pedestrians and cyclists to motorized traffic is negatively associated with vehicle-injury counts. This indicates that as pedestrians' and cyclists' activities increase, vehicle-injury frequencies decrease likely due to an increase in drivers' level of concentration and a decrease in operating speed.

Similar to the vehicle-injury dataset, the Dirichlet precision parameter k is close to 0 in the grade crossing dataset, again suggesting that the underlying random intercept distribution is not normal. The posterior density of k based on both gamma and uniform priors is shown in Fig. 4. We have support for the specified gamma prior based on two model-fitting measures: the cross-validation predictive density and the predictive ability of the model in replicating excess zero values. The generalized linear Dirichlet process mixture model (the simple Poisson model with a Dirichlet mixture over the intercept) identifies around 8 (3, 18) latent components for the crossing dataset. Note that, since the grade crossing dataset is not highly over-dispersed, this is not an over-dispersed model in contrast to that used for the vehicle-injury dataset. The variance of the error term υ_ε was estimated to be very

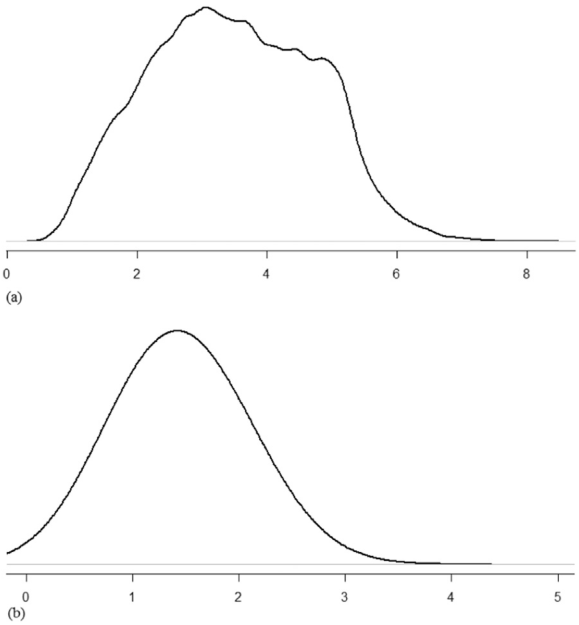

Fig. 4. Kernel posterior density plots of Dirichlet precision parameter k for the grade crossing dataset for two different prior densities: (a) k ~ uniform (0.3, 10); and (b) k ~ gamma (1, 1).

close to 0 when we analyzed the crossing dataset using the over-dispersed Dirichlet process mixture model, and so was dropped from further consideration.

It can be implied from Table 6 that the regression coefficients estimates obtained from different models are similar. Traffic exposure (the product of train flow and vehicle flow), train speed, interaction between the logarithm of road speed and the logarithm of the train flow to vehicle flow ratio were found to be positively associated with crash frequencies. In contrast, the presence of a gate in addition to the flashing lights and bells was found to reduce crash frequency. Finally, grade crossings located in Ontario, Quebec, Pacific region, and Atlantic region was found to have a lower chance of crash frequency compared to those located in the Prairie

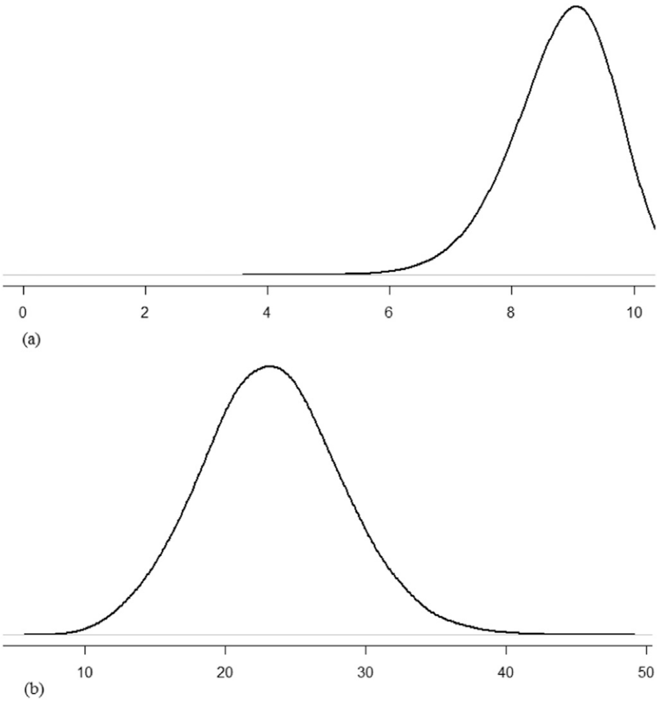

Fig. 5. Kernel posterior density plot of the precision parameter for the highway 401 dataset for two different prior densities: (a) k ~ uniform (0.3, 10); and (b) k ~ uniform (0.3, 100).

region. In terms of goodness-of-fit, the generalized linear Dirichlet process mixture model provides the highest log pseudo marginal likelihood; that is, −1687.65. This results in a log pseudo Bayes factor of 43.6 when comparing this model with the commonly used Poisson-gamma model (negative binomial), the conventional over-dispersed generalized linear model. When comparing the random intercept model with the Poisson-gamma model, a log pseudo Bayes factor of 20.96 provides support for the random intercept model in the grade crossing dataset.

We also examined whether the proposed Dirichlet process mixture model is able to properly generate such a large proportion of zero crashes observed in the grade crossings dataset. The results of the posterior predictive check in terms of the proportion of zero counts estimated a Bayesian p-value of 0.529, which is very close to the value 0.5 indicating a very good match between observed and replicated zero counts. Therefore, the Dirichlet process mixture model is excellent in this regard as well.

7.2. The multivariate setting

The results for the highway 401 dataset are reported in Table 7, where it can be seen that the Dirichlet precision parameter has a posterior mean of 8.752, away from the lower limit of 0 (Fig. 5). Therefore, the Dirichlet process mixture model does not appear to provide strong evidence for an underlying non-normal multivariate density for this dataset. To examine the sensitivity to the prior choice of a uniform (0.3, 10) distribution, for the precision parameter, we also analyzed the data using a uniform (0.3, 100) distribution. A kernel density plot of the precision parameter with different priors is shown in Fig. 5. Although the value of k varies, it remains bounded away from zero.

The log pseudo marginal likelihoods and coefficient estimates obtained from the Dirichlet process mixture multivariate model are similar to those from the standard multivariate Poisson-lognormal model. As traffic flow and segment length increase, crash frequencies of both type of severity increase. In contrast, an increase in median shoulder width or median width results in decreased injury-fatal crashes. The chance of property-damage-only crashes is higher among segments with a narrow median shoulder while the chance of injury-fatal crashes is lower among segments with paved outside shoulders. We also found that the degree of horizontal curve per km is negatively associated with both injury-fatal and property damage only crashes, and that two crash outcomes are highly correlated with a correlation of 0.876.

8. Conclusions

This study introduced Dirichlet process mixture models to analyze crash data in univariate and multivariate settings. The proposed technique derives from the Bayesian nonparametric literature, and presents a semiparametric model based on Dirichlet process priors. We followed Mukhopadhyay and Gelfand (1997) and Ohlssen et al. (2007) to refine the model to one which is not computationally cumbersome. The nonparametric part of the model manifests in the intercept in the univariate settings while it is found in the correlated random intercepts in the multivariate setting. Modeling intercepts nonparametrically allows us to conveniently retain the linear form of the vector of coefficients in relation to log-transformed responses (e.g., crash frequencies or differing injury-severity levels). This in turn retains usual interpretations made by conventional generalized linear models.

Using two simulated data, we first highlighted how the proposed model works and compares to conventional models used in road safety literature. We then adopted two real datasets for our univariate setting: (1) a vehicle-injury count dataset from signalized intersections that is somewhat highly over-dispersed and is characterized by a relatively large mean value; and (2) a railway grade crossing crash dataset that is characterized by the low mean value problem and excess zero counts. The results showed that Dirichlet process mixture models are applicable to different types of crash data. The proposed model allows us to examine the sensitivity to parametric assumptions, providing a better fit to both datasets compared to other conventional models commonly used in road safety studies.

In a multivariate setting, we used a highway segment data from Ontario to jointly model different crash types by severity. We showed how to extend the standard multivariate Poisson-lognormal model to a more flexible Dirichlet process mixture multivariate model, thereby accounting for dependence nonparametrically. We investigated the multivariate normal distribution assumption, and found it is reasonable, at least for this dataset. Our paper displays several advantages to Dirichlet process models. As Heydari et al. (2016) discuss, there are other advantages such as the ability to estimate the probability of similarities between pairs of observations.

Acknowledgments

This research was funded by the Natural Sciences and Engineering Research Council of Canada. We would like to thank Transport Canada (Rail Safety Directorate) for providing the railway grade crossing dataset, and the Ontario Ministry of Transportation for providing the highway 401 dataset. We wish to thank the data collection and preparation team at McGill University, in particular, Dr. Jillian Strauss, for the vehicle-injury dataset. We wish also to thank anonymous reviewers for their insightful comments and suggestions.

References

Anastasopoulos, P., Mannering, F., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153–159.

Anastasopoulos, P., Shankar, V.N., Haddock, J.E., Mannering, F., 2012. A multivariate Tobit analysis of highway accident-injury-severity rates. Accident Analysis and

Prevention 45, 110–119.
Anastasopoulos, P., 2016. Random parameters multivariate Tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis.. Analytic Methods in Accident Research 11, 17–32.
Anastasopoulos, P., Mannering, F., 2016. The effect of speed limits on drivers' choice of speed: a random parameters seemingly unrelated equations approach. Analytic Methods in Accident Research 10, 1–11.
Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to nonparametric problems. The Annals of Statistics 2 (6), 1152–1174.
Barua, S., El-Basyouny, K., Islam, T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1–15.
Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Analytic Methods in Accident Research 3–4, 56–91.
Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7 (4), 434–455.
Carlin, B.P., Louis, T.A., 2008. third editionBayesian Methods for Data Analysis. Chapman & Hall/CRC, Boca Raton.
Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Analytic Methods in Accident Research 3–4, 11–27.
Chen, E., Tarko, A., 2014. Modeling safety of highway work zones with random parameters and random effects models. Analytic Methods in Accident Research 1, 86–95.
Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with aggregated data: the random parameters negative binomial panel count data model. Analytic Methods in Accident Research 7, 37–49.
Dhavala, S.S., Datta, S., Mallick, B.K., Carroll, R.J., Khare, S., Lawhon, S.D., Adams, L.G., 2010. Bayesian modeling of MPSS data: gene expression analysis of bovine salmonella infection. Journal of the American Statistical Association 105 (491), 956–967.
Dorazio, R.M., 2009. On selecting a prior for the precision parameter of Dirichlet process mixture models. Journal of Statistical Planning and Inference 139 (9), 3384–3390.
El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41 (4), 820–828.
Escobar, M., West, M., 1998. Computing nonparametric hierarchical models. Practical Nonparametric and Semiparametric Bayesian Statistics 133, 1–22.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1 (2), 209–230.
Geedipally, S.R., Lord, D., Dhavala, S.S., 2014. A caution about using deviance information criterion while modelling traffic crashes. Safety Science 62, 495–498.
Gelfand, A., 1996. Model determination using sampling-based methods. In: Gilks, W., Richardson, S., Spiegelhalter, D. (Eds.), Markov Chain Monte Carlo in Practice. Chapman & Hall, Suffolk.
Gelfand, A.E., Dey, D.K., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling-based methods (with discussion) Bayesian Statistics 4. Clarendon, Oxford, 147–169.
Gelfand, A., Kottas, A., 2002. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. Journal of Computational and Graphical Statistics 11 (2), 289–305.
Gelman, A., Meng, X.L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733–807.
Gelman, A., Rubin, D., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7 (4), 457–511.
Gershman, S.J., Blei, D.M., 2012. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology 56 (1), 1–12.
Hauer, E., 1997. Observational Before-After Studies in Road Safety. Elsevier Science Ltd. Oxford, United Kingdom.
Heydari, S., Fu, L., 2015. Developing safety performance functions for railway grade crossings: a case study of Canada. 2015 Joint Rail Conference, JRC 2015, http://dx.doi.org/10.1115/JRC2015-5768.
Heydari, S., Fu, L., Lord, D., Mallick, B.K., 2016. Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. Analytic Methods in Accident Research 9, 27–43.
Heydari, S., Miranda-Moreno, L.F., Lord, D., Fu, L., 2014. Bayesian methodology to estimate and update safety performance functions under limited data conditions: a sensitivity analysis. Accident Analysis and Prevention 64, 41–51.
Hjort, N., Holmes, C., Müller, P., Walker, S.G., 2010. Bayesian Nonparametrics: Principles and Practice. Cambridge University Press.
Ishwaran, H., 2000. Inference for the random effects in Bayesian generalized linear mixed models. ASA Proceedings of the Bayesian Statistical Science Section, 1–10. Available at ⟨http://www.bio.ri.ccf.org/Resume/Pages/Ishwaran/publications.html⟩
Ishwaran, H., James, L.F., 2001. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96 (453), 161–173.
Jara, A., Garcia-Zattera, M.J., Lesaffre, E., 2007. A Dirichlet process mixture model for the analysis of correlated binary responses. Computational Statistics and Data Analysis 51 (11), 5402–5415.
Karlaftis, M., Tarko, A., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention 30 (4), 425–433.
Kim, D., Washington, S., 2006. The significance of endogeneity problem in crash models: an examination of left-turn lanes in intersection crash models. Accident Analysis and Prevention 38 (6), 1094–1100.
Kleinman, K.P., Ibrahim, J.G., 1998. A semiparametric Bayesian approach to the random effects model. Biometrics 54 (3), 921–938.
Lee, J., Abdel-Aty, M., Jiang, X., 2015. Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis and Prevention 78, 146–154.
Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10, 325–337.
Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964–975.
Mannering, F., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research 1, 1–22.
Mannering, F., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1–16.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models 2nd ed. Chapman & Hall, London, England.
Milton, J.C., Mannering, F.L., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25 (4), 395–413.
Mothafer, G.I.M.A., Yamamoto, T., Shankar, V.N., 2016. Evaluating crash type covariances and roadway geometric marginal effects using the multivariate Poisson gamma mixture model. Analytic Methods in Accident Research 9, 16–26.
Mukhopadhyay, S., Gelfand, A.E., 1997. Dirichlet process mixed generalized linear models. Journal of the American Statistical Association 92 (438), 633–639.
Muller, P., Quintana, F.A., 2004. Nonparametric Bayesian data analysis. Statistical Science 19 (1), 95–110.
Muller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83 (1), 67–79.
Muller, P., Quintana, F.A., Rosner, G.L., 2007. Semiparametric Bayesian inference for multilevel repeated measurement data. Biometrics 63 (1), 280–289.
Murugiah, S., Sweeting, T., 2012. Selecting the precision parameter prior in Dirichlet process mixture models. Journal of Statistical Planning and Inference 142 (7), 1947–1959.
Neal, R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational Statistics 9 (2), 249–265.
Ntzoufras, I., 2009. Bayesian Modeling Using WinBUGS. John Wiley & Sons.
Ohlssen, D.I., Sharples, L.D., Spiegelhalter, D.J., 2007. Flexible random-effects models using Bayesian semi-parametric models: application to institutional comparisons. Statistics in Medicine 26 (9), 2088–2112.
Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683–691.
Persaud, B.P., 1994. Accident prediction models for rural roads. Canadian Journal of Civil Engineering 21 (4), 547–554.

Rubin, D.B., 1984. Bayesian justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics 12 (4), 1151–1172.

Serhiyenko, V., Mamun, S.A., Ivan, J.N., Ravishankar, N., 2016. Fast Bayesian inference for modeling multivariate crash counts. Analytic Methods in Accident Research 9, 44–53.

Shaheed, M., Grikitza, K., 2014. A latent class analysis of single-vehicle motorcycle crash severity outcomes. Analytic Methods in Accident Research 2, 30–38.

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41 (7), 627–640.

Shirazi, M., Lord, D., Dhaval, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: characteristics and applications to crash data. Accident Analysis and Prevention 91, 10–18.

Strauss, G., Miranda-Moreno, L.F., Morency, P., 2014. Multimodal injury risk analysis of road users at signalized and non-signalized intersections. Accident Analysis and Prevention 71, 201–209.

Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. Safety Science 50 (4), 1141–1151.

Venkataraman, N., Shankar, V., Ulfarsson, G., Deptuch, D., 2014. A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies. Analytic Methods in Accident Research 2, 12–20.

Vehtari, A., Lampinen, J., 2002. Bayesian model assessment and comparison using cross-validation predictive densities. Neural Computation 14 (10), 2439–2468.

Walker, S.G., Damien, P., Laud, P.W., Smith, A.F.M., 1999. Bayesian nonparametric inference for random distributions and related functions (with discussion). Journal of the Royal Statistical Society, Series B 61 (3), 485–527.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and Econometric Methods for Transportation Data Analysis, Second Edition. Chapman Hall/CRC, Boca Raton, FL.

Wu, Z., Sharma, A., Mannering, F., Wang, S., 2013. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. Accident Analysis and Prevention 54, 90–98.

Xiong, Y., Mannering, F., 2013. The heteroscedastic effects of guardian supervision on adolescent driver-injury severities: a finite mixture-random parameters approach. Transportation Research Part B 49, 39–54.

Yu, R., Wang, X., Yang, K., Abdel-Aty, M., 2016. Crash risk analysis for Shanghai urban expressways: a Bayesian semi-parametric modeling approach. Accident Analysis and Prevention 95, 495–502.

Zhan, X., Aziz, H.M.A., Ukkusuri, S.V., 2015. An efficient parallel sampling technique for multivariate Poisson-lognormal model: analysis with two crash count datasets. Analytic Methods in Accident Research 8, 45–60.

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

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Research 1, 39–52.