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[b,c], Lawrence Jopseph[d], Luis F. Miranda-Moreno[e]

[a] *Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue W., Waterloo, Ontario, Canada N2L 3G1*
[b] *Department of Civil & Environmental Engineering, University of Waterloo, Waterloo, ON, Canada, N2L 3G1*
[c] *Intelligent Transportation Systems Research Center, Wuhan University of Technology, Mailbox 125, No. 1040 Heping Road, Wuhan, Hubei 430063, China*
[d] *Department of Biostatistics and Epidemiology, McGill University, 687 Pine Avenue West, Montreal, Canada*
[e] *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, 1992) 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. Jopseph), 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

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 pre-specified 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, 2008Venkataraman et al., 2014; Wu et al., 2013; Chen and Tarko, 2014; Mannering and Bhat, 2014; Barua et al., 2015; Coruh et al., 2015). 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., in press). 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.