Analytic Methods in Accident Research 9 (2016) 27–43

Contents lists available at ScienceDirect

Analytic Methods in Accident Research

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

Multilevel Dirichlet process mixture analysis of railway grade crossing crash data

Shahram Heydari [a,*], Liping Fu [a,1], Dominique Lord [b,2], Bani K. Mallick [c,3]

[a] Department of Civil and Environmental Engineering, University of Waterloo, 200 University Avenue W., Ontario, Canada N2L 3G1
[b] Zachary Department of Civil Engineering, Texas A&M University, College Station, TX, USA
[c] Department of Statistics, Texas A&M University, College Station, TX, USA

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form
5 February 2016
Accepted 5 February 2016
Available online 27 February 2016

Keywords:
Unobserved heterogeneity
Latent subpopulations
Random effects models
Finite mixture models
Spatial/regional multilevel models
Dirichlet process mixture models

ABSTRACT

This article introduces a flexible Bayesian semiparametric approach to analyzing crash data that are of hierarchical or multilevel nature. We extend the traditional varying intercept (random effects) multilevel model by relaxing its standard parametric distributional assumption. While accounting for unobserved cross-group heterogeneity in the data through intercept, the proposed method allows identifying latent subpopulations (and consequently outliers) in data based on a Dirichlet process mixture. It also allows estimating the number of latent subpopulations using an elegant mathematical structure instead of prespecifying this number arbitrarily as in conventional latent class or finite mixture models. In this paper, we evaluate our method on two recent railway grade crossing crash datasets, at province and municipality levels, from Canada for the years 2008–2013. We use cross-validation predictive densities and pseudo-Bayes factor for Bayesian model selection. While confirming the need for a multilevel modeling approach for both datasets, the results reveal the inadequacy of the standard parametric assumption in the varying intercept model for the municipality-level dataset. In fact, our proposed method is shown to improve model fitting significantly for the latter data. In a fully probabilistic framework, we also identify the expected number of latent clusters that share similar unidentified features among Canadian provinces and municipalities. It is possible thus to further investigate the reasons for such similarities and dissimilarities. This can have important policy implications for various safety management programs.

1. Introduction

Crash data are often characterized by a multilevel (hierarchical) structure in which observations at the lower level(s) of the hierarchy are nested in different groups (e.g., vehicles, sites, geographical areas, etc.) at the higher level(s) (Huang and Abdel-Aty, 2010; Dupont et al., 2013). Due to unobserved group-specific factors, such a hierarchical structure challenges the basic assumption of independent residuals since observations nested in the same groups usually share similar unknown

[*] Corresponding author. Tel.: +1 514 398 6589.
E-mail addresses: shahram.heydari@uwaterloo.ca (S. Heydari), lfu@uwaterloo.ca (L. Fu), d-lord@tamu.edu (D. Lord), bmallick@stat.tamu.edu (B.K. Mallick).
[1] Tel.: +1 519 888 4567.
[2] Tel.: +1 979 458 3949.
[3] Tel.: +1 979 845 1275.

http://dx.doi.org/10.1016/j.amar.2016.02.001
2213-6657/© 2016 Elsevier Ltd. All rights reserved.

and/or unmeasured traits and are thus correlated (Heydari et al., 2014a). In fact, if the hierarchical structure of the data is not accounted for through adequate statistical techniques, the estimated standard errors could be underestimated, resulting in erroneously estimated narrow confidence intervals (Lenguerrand et al., 2006; Dupont et al., 2013). Given the importance of the problem, instances of multilevel modeling in road safety have been numerous over the last decade; see, for example, Jones and Jørgensen (2003), Kim et al. (2007), Yannis et al. (2007), Huang et al. (2008), Helai et al. (2008), Cruzado and Donnell (2010), Heydari et al. (2014a) and Islam and El-Basyouny (2015). Readers are referred to Huang and Abdel-Aty (2010) and Dupont et al. (2013) for a comprehensive review of multilevel modeling in road safety literature.

In road safety, the multilevel structure of the data is often due to the nesting of observations in various geographical areas (Yannis et al., 2007, 2008; Huang and Abdel-Aty, 2010; Dupont et al. 2013; Papadimitriou et al., 2014). In such circumstances, it is quite plausible to speculate that sites such as railway grade crossings situated in the same regions share some similar unknown characteristics. For instance, these characteristics can be generated as a result of regional traffic regulations, driver demography and behavior, climate-related features, etc. Therefore, spatial dependencies may exist among sites sampled from similar geographical areas. In this regard, for example, Papadimitriou et al. (2014) investigated motorcycle riding under the influence of alcohol in 19 European countries and found significant regional variations.

With respect to the spatial concept, it should be noted that the conditional autoregressive model incorporating structured spatial random effects is one of the major spatial models used in road safety literature (Aguero-Valverde, 2013; Wang and Kockelman, 2013; Barua et al., 2014). It is important to highlight that the conditional autoregressive model does not differentiate between separate geographical areas, whereas it estimates spatial random effects (neighborhood effects) to account for the proximity of sites (e.g., intersections) that might share similar unobserved covariates (Aguero-Valverde, 2013; Dupont et al., 2013). For that reason, when the interest is in explicitly modeling the effect of geographical areas (or separation of geographical areas), as in this paper, the multilevel framework is a viable technique to accommodate spatial dependencies in the analysis (Huang and Abdel-Aty, 2010; Dupont et al. 2013).

In multilevel data, as discussed earlier, it is essential to account for group-specific effects. Three main approaches have been proposed in the literature to address this need: random effects models, random parameters models, and latent class or finite mixture models. Random effects models assume fixed parameters associated with the covariates but varying intercept or error term (Kim et al., 2007; Heydari et al., 2014a). In contrast to random effects models, in multilevel settings, random parameters models allow model covariates to vary across groups of observations to account for cross-group heterogeneity in data (Yannis et al., 2008; Islam and El-Basyouny, 2015). In general, random parameters models constitute therefore a more comprehensive way of overcoming unobserved heterogeneity in crash data including multilevel crash data, in comparison to random effects models. The higher quality and performance of random parameters models obviously comes with a higher cost in terms of computational complexity compared to random effects models (Chen and Tarko, 2014; Venkataraman et al., 2014). For a discussion related to random effects models and random parameters models, see Anastasopoulos and Mannering (2009), Lord and Mannering (2010), and Chen and Tarko (2014).

The finite mixture modeling approach (Park and Lord, 2009; Zou et al., 2014; Xiong and Mannering, 2013) is another alternative to overcome unobserved heterogeneity in crash data. However, to our knowledge, the application of finite mixture models in multilevel traffic safety studies has been limited in contrast to single-level safety studies. For a comparison between random parameters models and finite mixtures or latent class models, interested readers are referred to Behnood et al. (2014) and Mannering and Bhat (2014). Due to the higher computational complexity involved in random parameters models and finite mixture models, the majority of those studies involving multilevel analyses have used random effects models (Vanlaar, 2005, Lenguerrand et al., 2006; Kim et al., 2007; Helai et al., 2008; Park et al., 2010; Yannis et al., 2010; Jovanis et al., 2011; Papadimitriou et al., 2014). In this paper, among other reasons, we therefore focus on the use of random effects models (in particular varying intercept models) in multilevel settings. We discuss the limitations associated with random effects models and provide a flexible latent class model to circumvent such limitations.

To clarify one problem that may arise when adopting standard random effects models, suppose a multilevel scenario in which the modeler is only interested in potential variations in intercept (varying intercept model) among groups (e.g., geographical areas). A simplistic approach is to assume that all groups have exactly the same intercept and that there is no extra variability due to grouping in data. Obviously, this assumption does not take into consideration the fact that there might be some unknown and/or unmeasured attributes that change between groups. Basically, this approach ignores the hierarchical structure of the data.

In the aforementioned scenario, two major approaches have been proposed in the literature to account for group-specific effects and tackling unobserved heterogeneity through intercept. The first approach is estimating the intercepts for the individual groups separately based on the belief that they differ completely from each other, that is, the assumption of complete independence (Ohlssen et al., 2007). This assumption is not realistic since groups of observations (e.g., intersections or municipalities) are not totally dissimilar and they certainly share some similar features. A more appropriate approach, which is also the most commonly applied, is to assume that intercepts vary between groups but are generated from the same population. Thus, intercepts are assumed to share a common distribution being usually a unimodal normal distribution; i.e., a standard distributional assumption. Depending on the extent to which standard distributional assumptions are capable of capturing heterogeneity in a given data, say, in the form of random intercepts, the results would be biased by various degrees. It should be noted that standard distributional assumptions in traditional random effects models – such as normally distributed random effects models – usually do not accommodate skewness, kurtosis, and multimodality (Xiong and Mannering, 2013).