A Multivariate Heterogeneous-Dispersion Count Model for Asymmetric Interdependent Freeway Crash Types

Ghasak I.M.A. Mothafer

Department of Civil Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
Tel: +81-52-789-3565 Fax: +81-52-789-5728
Email: mothafer.g.i.m.a@c.mbox.nagoya-u.ac.jp

Toshiyuki Yamamoto

Institute of Materials and Systems for Sustainability, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
Tel: +81-52-789-4636 Fax: +81-52-789-5728
Email: yamamoto@civil.nagoya-u.ac.jp

Venkataraman N. Shankar**

Department of Civil, Environmental and Construction Engineering
Box 41023
Texas Tech University, Lubbock, TX 79409
Tel: 806-742-3189
Email: venky.shankar@ttu.edu

** Corresponding author

Revision submitted for publication consideration to Transportation Research Part B

Nov. 30, 2017

ABSTRACT

A multivariate count model is developed by introducing a simple and practical formula. The formulation begins with a modification of the standard ordered response model to adopt the count outcomes nature. This modification is accomplished by introducing a non-linear asymmetric interdependence structure among the error terms using the copula-based model. To avoid simulation maximum-likelihood for evaluating the multi-outcome density, we utilize the composite marginal likelihood (CML) approach. The proposed copula-based model with the CML approach allows for asymmetric (tail) dependency without a need for a simulation mechanism. Non-parametric graphical techniques with the empirical copula as well as conventional goodness-of-fit statistics are utilized to guide copula selection. In addition, unobserved heterogeneity across observations is also addressed through a heterogeneous dispersion parameter in the proposed model. The heterogeneous dispersion parameter model is a suitable alternative to random parameter count models in that captures heterogeneity in variance, while allowing for closed form while the latter needs numerical integration or simulation.

We apply these techniques to study the interdependence structure among four types of traffic crashes using three years (2005-2007) of cross-sectional crash data record for 274 multilane freeway segments in the State of Washington, USA. These four categories of crash types are the rear end; sideswipe; fixed objects and other crash types. The empirical results show a significant presence of unobserved heterogeneous dependency across these types of crashes. The results indicate the important role of unobserved heterogeneity in variance and covariance structure estimation. An important outcome of this result is that it can affect inference on the relative impact of roadway geometrics on crash occurrence. For example, we find that horizontal curve related parameters on freeway segments substantially increase the joint likelihood of rear-end, sideswipe, fixed objects and other crash types, when compared to the characteristics of vertical curves.

Keywords: Multivariate count data; copula; composite marginal likelihood; crash type correlations; variance-covariance structure; heterogeneous dispersion

1. INTRODUCTION

1.1 Background

The concept of multivariate count data modeling appears in many econometric applications. Multivariate count data modeling arises from the need for predicting the probability of several random integer non-negative outcomes simultaneously. This concept offers a better understanding of the interdependence of several random outcomes. The state of the art in estimating the interdependence of multiple traffic safety outcomes involves simulation based parameter estimation. One recent exception to this approach is the work of Bhat et al. (2014a) who have addressed three major types of multivariate count data approaches regarding the econometric formulation structure. The authors proposed a seminal perspective along three tracks of thought: a) via a general multivariate count model for obtaining the joint probability (usually in non-closed form); b) via a combination of a discrete-continuous data model in which count data are treated as random outcomes; and c) via a joint discrete choice-count model that accounts for the utility of discrete events.

In the first category, namely, multivariate count models, typically, there are five multivariate count models which offer a correlation structure among frequencies of the random outcomes: Multivariate Poisson model; multivariate negative binomial model; multivariate Poisson-gamma mixture model; multivariate Poisson-log-normal model and latent Poisson-normal model (Winkelmann, 2008). In the current paper, this approach is adopted to address a joint probability distribution that ties the random count outcomes through structural error terms (random unobserved heterogeneity) using the latent Poisson-normal model. Correlated counts in this model are explained as a realization of an underlying (latent) continuous random variable. Van Ophem (1999) and other studies (Castro et al., 2012; Narayanamoorthy et al., 2013; Yamamoto and Morikawa, 2013 and Bhat et al., 2014b; Bhat et al., 2015; Bhat et al., 2016a; Bhat et al., 2016b; Lavieri et al., 2017 and Bhat et al., 2017) utilized this model with the main assumption that the error term component is mapped from a normal distribution. The above-mentioned studies parameterized the threshold of a generalized ordered response (GOR) model as a function of a count distribution. The advantage of this model lies in its flexibility to handle both positive and negative dependency structure among the error terms. The flexibility in dependency is particularly useful in traffic crash analysis since the dependency might vary by context due to the nature of the

unobserved heterogeneity (for example, rural versus urban interstates; environmental heterogeneous contexts such as high-rain versus high-snow segments; and recreational versus commuting corridors, see Lavieri and Bhat 2016 and Bhat et al., 2016c). Mannering et al. (2016) briefly describe approaches to account for multivariate outcome modeling in the presence of unobserved heterogeneity. The authors stress the need for flexible correlation models that are unrestrictive on the nature of the dependency among outcomes. To address this need, we take an approach to include a non-linear asymmetric distribution dependency structure by adopting a copula-based concept. A copula is a tool to generate a multivariate distribution from univariate marginals (see for example Bhat and Eluru, 2009). Therefore, two steps are usually involved in the development of a copula: a) identifying the marginals, and b) determining the appropriate copula for accommodating the dependence structure. (So, the copula can be seen as a link between the marginals and the joint cumulative distribution. However, for discrete random variables, it must be noted that the associated copula representation may not be unique).

In the modeling of traffic crash count data, Poisson or negative binomial (NB) distributions are typically used as marginal distributions. However, as opposed to the usual bivariate case, accommodating the dependence structure in a multivariate case through the use of dependence parameters for each pair of marginals remains a challenge. The published literature suggests two approaches. The first approach involves the use of the mixture of powers concept (MOP) with some restrictions (see Zimmer and Trivedi, 2006; and Nikoloulopoulos and Karlis, 2010; Shi and Valdex, 2014).[1] Lee (1983) provided a normal copula through a transformation of non-normal disturbances, so that trivariate marginals can accommodate three parameters of dependency; however, this occurred at the expense of a closed-form. Hüsler and Reiss (1989) and Joe (1999) employ multivariate copulas with adequate dependence parameters, but in their approach they used a multivariate normal distribution only with a need for numerical integration. The second approach adopts the composite marginal likelihood (CML) technique. The CML approach has been used to overcome multi-dimensional complex dependencies without a need to evaluate the full likelihood function (see Ferdous et al., 2010; Sener et al., 2010; Castro et al., 2012, 2013; Paleti and Bhat, 2013; Yamamoto and Morikawa; 2013; Bhat et al., 2014c; Bhat and Dubey, 2014; Bhat et al., 2015 and Bhat et al.,2017). The CML approach is rooted in a general class of composite likelihood

[1] In the case of quadrivariate count outcomes the MOP approach produces three dependence parameters, so there are (I-1) parameter estimates of (I) count outcomes.

approaches (Lindsay, 1988). Both the MOP and CML approaches avoid using the simulation maximum likelihood method to evaluate the multivariate density of the count outcomes problem.

With regard to applications of count models in traffic crash analysis, a substantial body of literature exists (see Lord and Mannering, 2010; Mannering and Bhat, 2014; Mannering et al., 2016 for an exhaustive review). While some of the studies simultaneously considered crash frequency and severity (Chiou et al., 2013; Ye et al., 2013) the literature on the simultaneous treatment of multiple crash types dates back to Ma et al. (2008) and Park and Lord (2007). Other recent examples employing a Bayesian multivariate approach include Aguero et al. (2009), El-Basyouny and Sayed (2009), Imprialou et al. (2015), Lee et al. (2015), Li et al. (2015), Heydari et al. (2017) and Cheng et al. (2017). Two approaches are suggested to tackle the computational time problem of the fully Bayesian fitting with Markov Chain Monte Carlo (MCMC) when constructing the multivariate Poisson lognormal model. These approaches are the parallel sampling scheme by Zhan et al. (2015) and the integrate nested Laplace approximation (INLA) by Wang et al. (2017). Dong et al. (2014) have used a multivariate random-parameters zero-inflated negative binomial regression model to estimate crash frequencies of different types at intersections. Anastasopoulos et al. (2014) and Zeng et al. (2017) evaluated crash rates instead of crash frequencies by using the multivariate Tobit model to analyze the severity level on the freeway. In the injury analysis domain, in particular, Rana et al. (2010) used copula-based approach for addressing endogeneity in models of severity of traffic crash injuries while Yasmin et al. (2014) have used the same approach to examine driver injury severity in two vehicle crashes.

Nashad et al. (2016) have used the copula approach to investigate pedestrian and bicycle crashes. In their work, they utilized the negative binomial model as a marginal distribution to construct the copula-based model. The authors proposed several functions to parameterize the copula correlation parameter, rather than parameterizing the expected crash count. Their developed model is suitable for two dimensions only, although suggest potential extensions to higher dimensions.

In the context of the state-of-the-art econometric models, random parameters models have seen significant use in the address of unobserved heterogeneity among different observed individuals (in our case, segments). Generally, random parameter count models provides richer inferences compared to the classical fixed parameter (Mannering et al., 2016; Anastasopoulos and Mannering, 2009, 2011, 2016; Dong et al., 2014; Venkataraman et al., 2014; Coruh et al., 2015;

Barua et al., 2016; Bhat et al., 2017; Bhat and Lavieri, 2017). The estimation procedure for random parameters models involving simulation precludes their use in a variety of situations where closed form approaches are sufficient.

1.2 Aims of this study

The aim of this paper is to contribute to the literature on multivariate cross sectional count models, in particular, the efficient estimation of parameters while addressing unobserved heterogeneity across segments. The efforts in this paper are motivated by the recent work of Castro et al. (2012, 2013) and Bhat et al. (2014c, 2017). A characteristic of these studies is that the discrete ordered response approach was used. Given the questions that remain from these studies, namely the practicality of estimation of multidimensional outcomes under the presence of flexible dependencies, we apply a simple and practical approach. This approach involves a copula-based formulation using CML estimation to model dependence across the observational units. This approach provides flexibility in handling multiple marginals while taking advantage of the CML technique.

Our contribution to the literature is to provide for a rigorous approach to investigate dependency across crash types (including asymmetric tail) due to unobserved heterogeneity, without the need for numerical simulation of the likelihood. We demonstrate the contributions through a sequence of analyses – we first show the benchmarking of various copula alternatives for crash types that identifies the Gumbel copula as the preferred copula; we then show that the heterogeneous dispersion parameter count model structure is statistically preferable to a random parameter structure for the crash dataset used in this study; and finally, we then use the heterogeneous dispersion count model structure in a multivariate Gumbel copula formulation by accommodating the influence of geometrics and traffic volume on the Gumbel tail dependency parameter.

We then present a comprehensive discussion of the performance of the Gumbel copula (baselined against the independent copula) with respect to the dependency parameter variation by crash type pairs; in terms of the covariances of crash type pairs; and in terms of the variance in total crash count captured by the Gumbel copula. The Gumbel copula is then used to evaluate segment level correlations of crash type pairs as well marginal effects of the statistically significant geometric and traffic volume variables.

The rest of this paper is structured as follows. The next section provides the building blocks of the model in terms of formulation and inference. Section three describes the dataset including the selection of the crash types and the explanatory variables. Section four illustrates an application of the proposed crash type copula model and comparisons with the empirical copula. The final section summarizes the important findings and conclusions from the study.

2. METHODOLOGY

In this section, we begin with the formulation of a multivariate constant-dispersion copula-based count model (MHOCC) of crash types. The basis is a generalized ordered response (GOR) model in which a symmetrical interdependence among the error terms of crash count types is assumed. Next, a non-linear flexible correlation structure is introduced using the copula function. The MHOCC model assumes the dispersion parameter in the negative binomial marginal distribution is a constant among all observed individuals (segments). This is extended later to our final model, multivariate heteroscedastic copula-based count model (MHECC), which is capable to accommodate the heterogeneity effect among the observed specific segments for each given crash type.

2.1 Multivariate constant-dispersion copula-based count model

2.1.1 Ordered response model with count data

Following the generalized ordered model representation in Castro et al. (2012) [2], we assume q $(q = 1,2,...,Q)$ to represent the number of segments (or observation units), and i $(i = 1,2,...,I)$ to be the index of the crash type. Assume a count crash type variable y_{iq} can take the values k_{iq}, where $k_{iq} = 0,1,2,...$ is a stochastic count crash number of a specific type of crash i on a specific interstate segment q. Assuming a latent variable y_{iq}^{*} corresponding to the latent propensity underlying the observed count variable of y_{iq}, we can write:

$$y_{iq}^{*} = \boldsymbol{\varpi}_i^{T}\mathbf{z}_{iq} + \varepsilon_{iq}, \qquad y_{iq} = k_{iq} \quad \text{if} \quad \eta_i^{k_{iq}-1} < y_{iq}^{*} < \eta_i^{k_{iq}} \tag{1}$$

[2] Recently, Nashad et al., (2016) derived copula functions for count models directly without using ordered response representation.

where $\mathbf{z}_{iq}$ is a ($L\times1$) vector of non-intercept explanatory variables which are associated with crash type *i*, $\boldsymbol{\varpi}_i^T$ is a ($L\times1$) column-vector of parameters. The latent variable y_{iq}^* is drawn from a univariate continuous distribution which is a normal distribution in the case of the ordered response probit model; y_{iq}^* is bounded by the thresholds $\eta_i^{k_{iq}-1}$ and $\eta_i^{k_{iq}}$ (thresholds follow the usual ordered response model cutpoint definitions); and ε_{iq} is an identically distributed error term across segments representing the unobserved heterogeneity influencing the latent propensity of a crash type. Since we deal with count data, let's assume that y_{iq} follows a discrete-count distribution like Poisson, negative binomial (NB), Poisson-lognormal or zero-inflated distribution. If we assume that $\boldsymbol{\varpi}_i^T = 0$, then we can write $y_{iq}^* = \varepsilon_{iq}$, which lead to,

$$\Pr\left(y_{iq}=k_{iq}\right)=\Pr\left(k_{iq}-1<y_{iq}\le k_{iq}\right)=\Pr\left(\eta_i^{k_{iq}-1}<\varepsilon_{iq}\le\eta_i^{k_{iq}}\right). \tag{2}$$

This relationship is essential to connect the continuous to the count distributions together. We can write Eq. (2) in terms of the cumulative density functions, as:

$$\begin{aligned}\Pr\left(y_{iq}\le k_{iq}\right)&=\Pr\left(\varepsilon_{iq}\le\eta_i^{k_{iq}}\right)\\&=F_i\left(k_{iq}\right)=H_i\left(\eta_i^{k_{iq}}\right)\end{aligned} \tag{3}$$

where F_i is a univariate cumulative density function of a count crash type variable y_{iq} (e.g., Poisson or negative binomial); and H_i is a univariate cumulative density function of a latent propensity of a crash type *i* (ε_{iq}) (e.g., normal or t-student). Then we can write,

$$\eta_i^{k_{iq}}=\begin{cases}-\infty & k_{iq}=-1\\ H_i^{-1}\left[\Pr\left(y_{iq}\le k_{iq}\right)\right] & k_{iq}=0,1,....\end{cases} \tag{4}$$

where H_i^{-1} is a univariate cumulative density inverse function. This relationship defines the threshold value $\eta_i^{k_{iq}}$ uniquely for any selected parametric marginal distribution $\Pr\left(y_{iq}\le k_{iq}\right)$ for a continuous marginal distribution, but not unique in case of the discrete-count distributions (see Nelson, 2013; Joe, 2014). The threshold now is not in a linear relationship as in the GOR model, instead, it follows the marginal distribution form. This configuration ensures of getting always a

positive number for the threshold, which is the essence of our developed count model[3]. At the moment, we can write the multivariate cumulative probability density function $\boldsymbol{H}_{(1,2,\ldots,I)}$ for a given segment q as:

$$\begin{aligned}\Pr\left(y_{1q} \le k_{1q}, y_{2q} \le k_{2q}, \ldots, y_{Iq} \le k_{Iq}\right) &= \Pr\left(\varepsilon_{1q} \le \eta_1^{k_{1q}}, \varepsilon_{2q} \le \eta_2^{k_{2q}}, \ldots, \varepsilon_{Iq} \le \eta_I^{k_{Iq}}\right) \\ &= H_{(1,2,\ldots,I)}\left(\eta_1^{k_{1q}}, \eta_2^{k_{2q}}, \ldots, \eta_I^{k_{Iq}}\right) \\ &= \int_{-\infty}^{\eta_1^{k_{1q}}} \int_{-\infty}^{\eta_2^{k_{2q}}} \ldots \int_{-\infty}^{\eta_I^{k_{Iq}}} f_{(1,2,\ldots,I)}\left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_I \mid \boldsymbol{\Theta}\right) d\varepsilon_1\, d\varepsilon_2 \ldots d\varepsilon_I\end{aligned} \tag{5}$$

where $f_{(1,2,\ldots,I)}$ is the multivariate probability density function of the I-dimensions, $\boldsymbol{\Theta}$ is the matrix of correlation among the error terms ε_{iq}. We can write the multivariate joint probability distribution function for a given segment q as:

$$\begin{aligned}&\Pr\left(y_{1q} = k_{1q}, y_{2q} = k_{2q}, \ldots, y_{Iq} = k_{Iq}\right) = \\ &\int_{\varepsilon_1=\eta_1^{k_{1q}-1}}^{\eta_1^{k_{1q}}} \int_{\varepsilon_2=\eta_2^{k_{2q}-1}}^{\eta_2^{k_{2q}}} \ldots \int_{\varepsilon_I=\eta_I^{k_{Iq}-1}}^{\eta_I^{k_{Iq}}} f_{(1,2,\ldots,I)}\left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_I \mid \boldsymbol{\Theta}\right) d\varepsilon_1\, d\varepsilon_2 .. d\varepsilon_I .\end{aligned} \tag{6}$$

Since Eq. (6) has no closed form, we evaluate it in copula form, which allows us to solve the integral of the joint distribution and to seek non-linear and asymmetric patterns of relationships among the error terms.

2.1.2 Copula with count data

Sklar's theorem (1959) provides that there exists a class of distribution function such that the n-dimensional cumulative distribution can be expressed in terms of the copula and the marginal. When y_{iq} are discrete (count) variables and F_i are discrete cdf's, the multivariate cumulative probability density function $H_{(1,2,\ldots,I)}$ for a given segment q (shown in Eq. (5)) can be constructed from $\left[F_1\left(y_{1q}\right), F_2\left(y_{2q}\right), \ldots, F_I\left(y_{Iq}\right)\right] \in Ran\left(F_1\right) \times Ran\left(F_2\right) \times \ldots \times Ran\left(F_I\right)$ where $Ran(F_i)$ denotes the

[3] In the context of parameterizing the threshold, Bhat et al. (2015, 2016a, 2016b) proposed to use the cumulative negative binomial model $F\left(k_{iq}\right)$ to accommodate the count outcome dimension using the $H_i^{-1}\left[\Pr\left(y_{iq} \le k_{iq}\right)\right] = \Phi_i^{-1}\left[F\left(k_{iq}\right)\right]$, where Φ_i^{-1} is the inverse function of the univariate cumulative standard normal distribution.

range of the marginals $F_i(.)$. Using the inverse cumulative density function approach via the multivariate copula $C_{(1,2,...,I)q}$, we can write:

$$\begin{aligned} & H_{(1,2,...,I)q}\left(\eta_1^{k_{1q}}, \eta_2^{k_{2q}}, ..., \eta_I^{k_{Iq}}\right) \\ & = H_{(1,2,...,I)q}\left\{H_1^{-1}\left[\Pr\left(y_{1q} \le k_{1q}\right)\right], H_2^{-1}\left[\Pr\left(y_{2q} \le k_{2q}\right)\right], ..., H_I^{-1}\left[\Pr\left(y_{Iq} \le k_{Iq}\right)\right]\right\} \\ & = C_{(1,2,...,I)q}\left[F_1\left(k_{1q}\right), F_2\left(k_{2q}\right), ..., F_I\left(k_{Iq}\right) | \boldsymbol{\theta}\right] \end{aligned} \tag{7}$$

For all $\left[F_1\left(y_{1q}\right), F_2\left(y_{2q}\right), ..., F_I\left(y_{Iq}\right)\right] \in [0,1]^I$; $\boldsymbol{\theta}$ is the matrix of correlation among the marginal distributions for a specified copula $\boldsymbol{\theta} = \begin{pmatrix} \theta_{11} & \cdots & \theta_{1I} \\ \vdots & \ddots & \vdots \\ \theta_{I1} & \cdots & \theta_{II} \end{pmatrix}$

by taking the derivative of both sides of Eq. (7) we can get:

$$\begin{aligned} & \frac{\partial H_{(1,2,..,I)q}\left(\varepsilon_{1q}, \varepsilon_{2q}, ..., \varepsilon_{Iq}\right)}{\partial \varepsilon_{1q}, \partial \varepsilon_{2q}, ..., \partial \varepsilon_{Iq}} = \frac{\partial C\left[F_1\left(y_{1q}\right), F_2\left(y_{2q}\right), ..., F_I\left(y_{Iq}\right)\right]}{\partial F_1\left(y_{1q}\right), \partial F_2\left(y_{2q}\right), ..., \partial F_I\left(y_{Iq}\right)} \\ & f_{(1,2,..,I)q}\left(\varepsilon_{1q}, \varepsilon_{2q}, ..., \varepsilon_{Iq} | \boldsymbol{\theta}\right) = c_{(1,2,..,I)}\left[F_1\left(y_{1q}\right), F_2\left(y_{2q}\right), ..., F_I\left(y_{Iq}\right)\right] . \prod_{i=1}^{I} f_i\left(y_{iq}\right) \end{aligned} \tag{8}$$

where $c_{(1,2,..,I)}$ is the multivariate copula density function, $f_i\left(y_{iq}\right)$ is the univariate density function of the marginal distribution i^{th}, now we can substitute the result of Eq. (8) into Eq. (6) and retrieve the integration boundaries using Eq. (4) to get:

$$\begin{aligned} & \Pr\left(y_{1q} = k_{1q}, y_{2q} = k_{2q}, ..., y_{Iq} = k_{Iq}\right) \\ & = \int_{F_1\left(k_{1q}-1\right)=H_1\left(\eta_1^{k_{1q}-1}\right)}^{F_1\left(k_{1q}\right)} \int_{F_2\left(k_{2q}-1\right)}^{F_2\left(k_{2q}\right)} .... \int_{F_I\left(k_{Iq}-1\right)}^{F_I\left(k_{Iq}\right)} c_{(1,2,..,I)}\left[F_1\left(y_{1q}\right), F_2\left(y_{2q}\right), ...\right. \\ & \left. ...., F_I\left(y_{Iq}\right) | \boldsymbol{\theta}\right] . \prod_{i=1}^{I} f_i\left(y_{iq}\right) dF_1\left(y_{1q}\right) dF_2\left(y_{2q}\right) ..... dF_I\left(y_{Iq}\right) \end{aligned} \tag{9}$$

Eq. (9) is the joint probability distribution of the multivariate crash count types written in terms of the copula density function. The copula approach offers an extensive range of parametric and non-parametric functions, but in general, it can be classified into two families. First, the elliptical copula, offers a non-closed form for Eq. (9) and the integral should be evaluated either numerically or by simulation. Second, the Archimedean copula offers the closed-form and can be evaluated by taking

differences of the copula function *C* for the same boundaries of Eq. (9). The model estimation is carried out after specifying a suitable marginal distribution *F* for the count outcome and an appropriate copula *C*. Discrete copulas are not unique; albeit this fact, it has been proven that such non-uniqueness is not a problem and the discrete count copula still inherits the dependency feature analogous to the continuous one. More details regarding this issue are presented by Denuit and Lambert (2005), Zimmer and Trividi (2006) and Genest and Nešlehová (2007).

2.2 Multivariate heterogeneous-dispersion copula count model

As the very basic count modeling, the count data can be modeled with a Poisson regression. The probability of a certain crash type count variable y_{iq} having k_{iq} accidents is shown as

$$P\left(y_{iq}=k_{iq}\mid\lambda_{iq}\right)=\frac{e^{-\lambda_{iq}}\times\lambda_{iq}^{k_{iq}}}{k_{iq}!} \tag{10}$$

where λ_{iq} is the parameter for crash type *i* of observed segment *q*. λ_{iq} is usually specified as a function of explanatory variables by log-linear function as $\lambda_{iq}=\exp\left(\boldsymbol{\beta}_i^T\ \mathbf{x}_{iq}\right)$ where $\mathbf{x}_{iq}$ is a $(L\times 1)$ vector of explanatory variables including constant that influence a certain type of crash *i* of segment *q* with corresponding $(L\times 1)$ set of parameter vector $\boldsymbol{\beta}_i$. Analogue to the conventional linear regression models, it is natural to add an error term ξ_{iq} to represent unobserved effect of omitted variables as $\lambda'_{iq}=\exp\left(\boldsymbol{\beta}_i^T\ \mathbf{x}_{iq}+\xi_{iq}\right)=\lambda_{iq}\nu_{iq}$ where $\nu_{iq}=\exp\left(\xi_{iq}\right)$. The unconditional probability is given as a result of the mixture probability function,

$$\Pr\left(y_{iq}=k_{iq}\middle|\lambda_{iq}\right)=\int_0^\infty P\left(y_{iq}=k_{iq}\mid\lambda_{iq},\nu_{iq}\right)f\left(\nu_{iq}\right)d\nu_{iq} \tag{11}$$

If ν_{iq} is assumed to follow log-normal distribution, Eq. (11) becomes Poisson-lognormal model. Unfortunately, Poisson-lognormal model doesn't have a closed form, so numerical integration or simulation is required. On the other hand, if ν_{iq} is assumed to follow a gamma distribution with $\nu_{iq}\overset{i.i.d.}{\sim}Gamma\left(\psi_{iq}^{-1},\psi_{iq}^{-1}\right)$, $E\left(\nu_{iq}\right)=1$ and $V_{iq}\left(\nu_{iq}\right)=\psi_{iq}$, a negative binomial type-II distribution (NBII) for each marginal distribution is a result of such conjugation between Poisson and gamma

distribution,

$$P_i\left(y_{iq}\right)=\frac{\Gamma\left(r+\psi_{iq}^{-1}\right)}{\Gamma\left(\psi_{iq}^{-1}\right)\Gamma\left(r+1\right)}\left(\frac{\psi_{iq}^{-1}}{\lambda_{iq}+\psi_{iq}^{-1}}\right)^{\psi_{iq}^{-1}}\left(\frac{\lambda_{iq}}{\lambda_{iq}+\psi_{iq}^{-1}}\right)^{r}. \tag{12}$$

The expected crash count is given as $E\left(y_{iq}\right)=\lambda_{iq}$ and variance $\Gamma_i\left(y_{iq}\right)=\lambda_{iq}+\psi_{iq}\left(\lambda_{iq}\right)^2$ (overdispersion occurs when ψ_{iq} > 0). For each observation the NBII *cdf* is obtained by summing the crash numbers from 0 to k_{iq} as:

$$\Pr\left(y_{iq}\le k_{iq}\,|\,\boldsymbol{\beta}_{\mathbf{i}}^T,\mathbf{x}_{iq}\right)=F_i\left(k_{iq}\right)=\sum_{r=0}^{k_{iq}}P_i\left(r\,|\,\boldsymbol{\beta}_{\mathbf{i}}^T,\mathbf{x}_{iq}\right). \tag{13}$$

Each marginal $F_i\left(k_{iq}\right)$ for each crash type *i* is determined conditionally on $\mathbf{x}_{iq}$, which is used to construct our copula model in Eq. (9) [4].

As mentioned in the introduction, unobserved heterogeneity across observations is also represented by the random parameter models (Mannering et al., 2016). If the coefficients are assumed to be randomly distributed, the coefficient vector can be represented by the sum of the fixed part $\boldsymbol{\beta}_i$ and the random part $\boldsymbol{\varphi}_i$. The Poisson parameter is now given as

$$\breve{\lambda}_{iq}=\exp\left[\left(\boldsymbol{\beta}_i^T+\boldsymbol{\varphi}_i^T\right)\mathbf{x}_{iq}+\xi_{iq}\right]. \tag{14}$$

If $\boldsymbol{\varphi}_i$ is assumed to follow normal distributions as usual, it requires numerical integration or simulation to calculate the unconditional probability regardless of the distributional assumption on ξ_{iq}. Eq. (14) can be also rewritten as

$$\breve{\lambda}_{iq}=\exp\left(\boldsymbol{\beta}_i^T\mathbf{x}_{iq}\right)\exp\left(\boldsymbol{\varphi}_i^T\mathbf{x}_{iq}+\xi_{iq}\right)=\lambda_{iq}\nu_{iq}^T. \tag{15}$$

where $\nu_{iq}^T=\exp\left(\boldsymbol{\varphi}_i^T\mathbf{x}_{iq}+\xi_{iq}\right)$. In the case that both $\boldsymbol{\varphi}_i$ and ξ_{iq} follow normal distributions, the sum also follows a normal distribution, then ν_{iq}^T follows a heterogeneous lognormal distribution

[4] Based on which distribution is selected to represent the $f\left(\nu_{iq}\right)$ term, there are seven appealing possible count models that appear in the literature. These models are the negative binomial type I, type II (the one we use), Poisson inverse Gaussian mixture, Poisson lognormal, Hurdle, Zero inflated, Finite mixture (Cameron and Trivedi, 2013).

depending on $\mathbf{x}_{iq}$. It means that the random coefficient Poisson-lognormal model is equivalent to the Poisson heterogeneous lognormal mixture model. As the gamma distribution's counterpart of the Poisson heterogeneous lognormal mixture model, we assume ν_{iq}^{T} follows a heterogeneous gamma distribution depending on $\mathbf{x}_{iq}$, then the resulting model becomes Poisson heterogeneous gamma mixture model, which can be also called as the heterogeneous negative binomial model. Parameterizing the dispersion parameter in the NBII marginal distribution of each crash count outcome does not need for a numerical/simulation solution. In our current model framework, considering the heterogeneity effect by structured dispersion parameter is expected to capture the unobserved heterogeneity as in the random parameter count models. The dispersion parameter in Eq. (12) becomes in that case,

$$\psi_{iq} = \gamma_i \exp\left(\mathbf{c}_i^{T}\boldsymbol{\delta}_{iq}\right) \tag{16}$$

where γ_i is a necessary constant to capture the dispersion in case no significant variables in $\boldsymbol{\delta}_{iq}$ which is a $(L\times 1)$ vector of non-intercept explanatory variables that are associated with each individual dispersion variable of given crash type *i*. $\mathbf{c}^{\mathrm{T}}$ is a $(L\times 1)$ column-vector of parameters to be estimated along with the others parameters. Since the homoskedastic name is implicit, the MHOCC model is obtained through suppressing Eq. (16) to carry constants only, in that case the variance is assumed to have the same size across all observations.

2.3 Choosing a copula function

As previously mentioned, several types of parametric copula functions (Archimedean and Elliptical) are available for model development. So far, there is no robust formula that assesses the goodness-of-fit of a copula without the need to investigate all the other types of copulas. Four graphical techniques are available to aid our selection of the parametric copula – these techniques give an initial insight to the dependency structure of the outcomes regardless of the marginal effect. These techniques are: a) the PP-plot which is most general, but least effective; b) tail dependence plot – general and more effective than the PP-plot; c) the K-plot which is used for Archimedean copulas only and finally, d) the t-plot which is most restrictive - useful for elliptical copulas but only for model diagnostic checking. We will select the first two graphical methods, the PP-plot and the tail dependence plot due to their generality for both Archimedean and Elliptical copulas.

The PP-plot for copula also known as the "Copula PP-plot" was introduced by Sun et al. (2008). The Copula PP-plot evaluates the probability values at each observation point corresponding to the theoretical copula function (parametric copula) and the empirical copula (non-parametric). The tail dependence-plot (Joe, 1997) focuses on visualizing the dependence of each parametric copula compared to the empirical copula for the upper and lower tails using the tail concentration function. The tail concentration function separates the dependencies into two parts (for two-dimension copula) which are upper and lower tails (Boucher et al., 2008). Suppose $Z \in [0,1]$, then we can write the tail concentration function as:

$$LR(Z) = \begin{cases} L(Z) & \text{if} \quad 0 \le Z < 0.5 \\ R(Z) & \text{if} \quad 0.5 \le Z \le 1 \end{cases} \tag{17}$$

given both, the lower $L(Z)$ and the upper $R(Z)$ tail functions are given as:

$$\begin{aligned} L(Z) &= \frac{\Pr\left(y_i \le F_i^{-1}(Z), y_j \le F_j^{-1}(Z)\right)}{\Pr\left(y_i \le F_i^{-1}(Z)\right)} = \frac{C(Z,Z)}{Z} \\ R(Z) &= \frac{\Pr\left(y_i > F_i^{-1}(Z), y_j > F_j^{-1}(Z)\right)}{\Pr\left(y_i > F_i^{-1}(Z)\right)} = \frac{C(1-Z,1-Z)}{1-Z} = \frac{1-2Z+C(Z,Z)}{1-Z} \end{aligned} \tag{18}$$

As implied above, we need to construct the empirical copula at first to conduct these graphical techniques. Let (m_{iq}, m_{jq}) be a pair of observed crash counts for types *i* and *j* on segment *q*. The bivariate empirical copula function $\tilde{C}^n_{(i,j)}$ (Deheuvels, 1979) is a function with a domain $\{a/Q : a = 0,1,.....,Q\}^2$ and marginals U_a and $V_b \in [0,1]$, which is formulated as:

$$\begin{aligned} \tilde{C}^n_{(i,j)}\left[\frac{a}{Q}, \frac{b}{Q}\right] &= \frac{\text{number of pairs}\left(m_{iq}, m_{jq}\right) \text{ with } m_{iq} \le m_{i(a)}, m_{jq} \le m_{j(b)}}{Q} \\ &= \frac{1}{Q}\sum_{q=1}^{Q} \mathrm{I}\left(\tilde{F}_i^n\left(m_{iq}\right) \le U_a, \tilde{F}_j^n\left(m_{jq}\right) \le V_b\right) \\ &= \frac{1}{Q}\sum_{q=1}^{Q} \mathrm{I}\left(R_i\left(m_{iq}\right) \le a, R_j\left(m_{jq}\right) \le b\right) \end{aligned} \tag{19}$$

where $m_{i(a)}$ and $m_{j(b)}, 1 \le a, b \le Q$ are order statistics from the sample,

$\widetilde{F}_i^n(m_{iq}) = \frac{1}{Q}\sum_{f=1}^{Q} \mathrm{I}(m_{if} \le m_{iq})$ and $\widetilde{F}_j^n(m_{jq}) = \frac{1}{Q}\sum_{f=1}^{Q} \mathrm{I}(m_{jf} \le m_{jq})$ are the empirical cumulative distribution functions of the observations, $R_i(m_{iq})$ and $R_j(m_{jq})$ are the rank functions[5] of the observed crash count in the dataset. $\mathrm{I}(.)$ denotes the indicator function that can take a value equal to 0 whenever its argument is false, and 1 otherwise. Table (1) lists available empirical copulas, in which 1/Q type is used in this study (see Hernandez-Maldonado et al., (2012) and Asquith (2016) for more details). The tails dependence of the empirical copula is constructed using Eq. (18).

2.4 Level of dependency of the copula function

The bivariate copula function includes one parameter which represents a measure of dependency between the marginal distributions. If the marginal distributions are independent, the level of dependency θ_{ij} would be equal to zero and the estimation could be carried out individually for each marginal. In general, it's not straightforward to interpret the level of dependency θ_{ij} like the case of Pearson product-moment correlation coefficient (except the case of the elliptical copula family), because of two reasons. First, the bivariate copula functions represent a non-linear relationship between the marginal distributions. Second, many of these copula functions don't require that $\theta_{ij} :\in [-1,+1]$, therefore, other non-parametric measures (like Kendall's 'τ' or Spearman's 'ρ') are commonly utilized instead (Cameron et al., 2004). In case of continuous y_{iq} variable is used, θ_{ij} is transformed to these measures, which are independent from the marginal distributions and bounded on the interval $[-1,+1]$. Marshall (1996), Bouyé et al. (2000) and Tajar et al. (2001) state that these measures of dependence are not useful in the case of discrete variables because θ_{ij} depends on the selection of marginal distributions, therefore extra care is required when interpreting θ_{ij} for count data. For this reason, we will maintain the same count marginal distribution for the same crash type along with the modeling processes to facilitate the comparison among the developed models.

---

[5] It can be seen that the rank function is expressed as $R_i(m_{iq}) = \sum_{f=1}^{Q} \mathrm{I}(m_{if} \le m_{iq})$ given that $\widetilde{F}_i^n(m_{iq}) = R_i(m_{iq})/Q$. Therefore, the empirical copula can be seen as the empirical distribution of the rank transformed data as shown in the last part of Eq. (19).

To keep the consistency with the heterogeneity concept in the MHECC model, we allowed the level of dependency to vary across all the observed individual segments. This idea has been accomplished by parameterizing θ_{ijq} as a function of a $\mathbf{o}_{ijq}$ $(L\times 1)$ vector of dependency influential pairwise variables among the observed crash type pair (*i, j*) for a given segment *q* as

$$\theta_{ijq}\left(\mathbf{o}_{ijq}\right)= f_{(i,j)}\left(\theta_{ij}+\mathbf{d}_{ij}^{T}\,\mathbf{o}_{ijq}\right) \tag{20}$$

where θ_{ij} is a necessary pairwise constant to reflect the correlation value in case no significant parameter in $\mathbf{o}_{ijq}$. $\mathbf{d}_{ij}^{T}$ is a $(L\times 1)$ column-vector without-intercept of parameters of a given crash type pair (*i, j*) to be estimated along with the others parameters. In similar way to the dispersion parameter function, the constant-dispersion MHOCC model is obtained through suppressing Eq. (20) to carry constants only, in that case the copula function is assumed fixed across all observations. The function selection of Eq.(20) depends on the range of the parameter θ_{ijq} of a given copula type to avoid any discontinuity in the copula function (see Bhat and Sener 2009; Bhat et al., 2010 and Nashad et al., 2016).

2.5 Composite marginal likelihood CML

The CML approach is useful for multi-dimensional dependencies as seen in Eq. (9) without a need to evaluate the full likelihood function. In this paper, we will use the pairwise marginal likelihood estimation method (see Ferdous et al., 2010; Sener et al., 2010; Castro et al., 2012, 2013; Paleti et al., 2013; Yamamoto and Morikawa, 2013 and Bhat et al., 2014c; Bhat and Dubey, 2014; Bhat et al., 2015; Bhat et al.,2016b and Bhat et al.,2017). The features of the bivariate copula can be obtained from Eq. (7) when I = 2 with the following properties: $C\left[F_1\left(y_{1q}\right),0\right]=C\left[0,F_2\left(y_{2q}\right)\right]=0$; $C\left[F_1\left(y_{1q}\right),1\right]=F_1\left(y_{1q}\right)$; $C\left[1,F_2\left(y_{2q}\right)\right]=F_2\left(y_{2q}\right)$. Let $\left(m_{1q},m_{2q},....,m_{Iq}\right)$ as the actual observed crash count of type *i* on a specific segment *q*. Let an index $\left(j=1,2,....,I\right)$, the full length parameter vector of the MHECC model is given as

$\boldsymbol{\zeta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_I; \gamma_1, \gamma_2, ..., \gamma_I; \mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_I; \mathbf{d}_{(1,1)}, \mathbf{d}_{(1,2)}, ..., \mathbf{d}_{(I,I-1)}\right)$ [6] and Eq. (9) collapses into I(I-1)/2 pairs of bivariate probability computations and it takes the form:

$$
\begin{aligned}
L_{CML_q}(\boldsymbol{\zeta}) &= \prod_{i=1}^{I-1}\prod_{j=i+1}^{I} \Pr\left(y_{iq} = m_{iq}, y_{jq} = m_{jq}\right) \\
&= \prod_{i=1}^{I-1}\prod_{j=i+1}^{I} \left\{ \int_{F_i(m_{iq}-1)}^{F_i(m_{iq})} \int_{F_j(m_{jq}-1)}^{F_j(m_{jq})} c_{(i,j)}\left[F_i(y_{iq}), F_j(y_{jq}) \mid \theta_{ij}\right] . \prod_{i=1}^{I} f_i(y_{iq})\, dF_i(y_{iq}) dF_j(y_{jq}) \right\} \\
&= \prod_{i=1}^{I-1}\prod_{j=i+1}^{I} \left\{ \begin{array}{l} C\left[F_i(m_{iq}), F_j(m_{jq}) \mid \theta_{ijq}\right] - C\left[F_i(m_{iq}-1), F_j(m_{jq}) \mid \theta_{ijq}\right] \\ -C\left[F_i(m_{iq}), F_j(m_{jq}-1) \mid \theta_{ijq}\right] + C\left[F_i(m_{iq}-1), F_j(m_{jq}-1) \mid \theta_{ijq}\right] \end{array} \right\}
\end{aligned}
\tag{21}
$$

where θ_{ijq} represents the level of dependency between the marginals $F_i(m_{iq}), F_j(m_{jq})$ for a certain copula function *C*. Eq. (21) is constructed now in a way to accommodate all types of variations in both the dispersion parameters and the level of dependency. The pairwise marginal likelihood across all segments can be computed using $L_{CML}(\boldsymbol{\zeta}) = \prod_{q=1}^{Q} L_{CML_q}(\boldsymbol{\zeta})$. The pairwise likelihood estimator $\hat{\boldsymbol{\zeta}}_{CML}$ is obtained by maximizing the logarithm of the $L_{CML}(\boldsymbol{\zeta})$ function with respect to the vector $\boldsymbol{\zeta}$, which is consistent, and asymptotically normal distributed with asymptotic mean $\boldsymbol{\zeta}$ and covariance matrix given by the inverse of Godambe's (1960) sandwich information matrix $\mathbf{G}(\boldsymbol{\zeta})$ (see Zhao and Joe, 2005; Ferdous et al. 2010 and Castro et al., 2012).

$$
\begin{aligned}
\mathbf{V}_{CML}(\boldsymbol{\zeta}) &= \left[\mathbf{G}(\boldsymbol{\zeta})\right]^{-1} \\
&= \left[\mathbf{H}(\boldsymbol{\zeta})\right]^{-1} \mathbf{J}(\boldsymbol{\zeta}) \left[\mathbf{H}(\boldsymbol{\zeta})\right]^{-1}, \text{where} \\
\mathbf{H}(\boldsymbol{\zeta}) &= E\left[-\frac{\partial^2 \ln L_{CMLq}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}\, \partial \boldsymbol{\zeta}'} \right] \text{and} \\
\mathbf{J}(\boldsymbol{\zeta}) &= E\left[\left(\frac{\partial \ln L_{CMLq}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}} \right) \left(\frac{\partial \ln L_{CMLq}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}'} \right) \right]
\end{aligned}
\tag{22}
$$

[6] Obviously, the dispersion parameter in Eq. (16), will be $\psi_i = \gamma_i$ and the level of dependency parameters will be fixed across all segments as θ_{ij} in Eq.(20), when the multivariate homoskedstic copula-based is selected, hence the vector will collapses into $\boldsymbol{\zeta} = \left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_I; \gamma_1, \gamma_2, ..., \gamma_I; \boldsymbol{\theta}\right)$

where $\mathbf{H}(\zeta)$ and $\mathbf{J}(\zeta)$ are Hessian and Jacobian matrices, respectively. Estimates of the Hessian and Jacobian matrices at the CML estimate ($\hat{\zeta}_{CML}$) as shown below are consistent with the results of the empirical copula:

$$\hat{\mathbf{H}}\left(\hat{\zeta}\right) = \frac{1}{Q}\sum_{q=1}^{Q}\left[-\frac{\partial^2 \ln L_{CMLq}(\zeta)}{\partial\zeta\,\partial\zeta'}\right]_{\hat{\zeta}}$$
$$= \left[\frac{1}{Q}\sum_{q=1}^{Q}\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\frac{\partial^2 \ln \Pr\left(y_{iq}=m_{iq}, y_{jq}=m_{jq}\right)}{\partial\zeta\,\partial\zeta'}\right]_{\hat{\zeta}}, \text{ and} \quad (23)$$
$$\hat{\mathbf{J}}\left(\hat{\zeta}\right) = \frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\frac{\partial \ln L_{CMLq}(\zeta)}{\partial\zeta}\right)\left(\frac{\partial \ln L_{CMLq}(\zeta)}{\partial\zeta'}\right)\right]_{\hat{\zeta}}.$$

2.6 Model selection

Nikoloulopoulos and Karlis (2009); Winkelmann (2012); Cameron and Trivedi (2013) utilized the Akaike information criterion (*AIC*) while Yasmin et al. (2014) utilized Schwarz Information Criterion (*BIC*) to select the copula that provides the best fit. The *BIC* performed better in large samples, whereas the *AIC* tends to be superior in small samples (Shumway and Stoffer, 2010). *AIC* and *BIC* criteria were implemented and the copula that provides the best fit is the one that correspond with the lowest values of these measures. The *AIC* and the *BIC* can be defined as follows: $AIC = -2\times\log(LL)+2\times(Q)$ and $BIC = -2\times\log(LL)+(\kappa)\times\log(Q)$, where κ is the number of parameters of the copula model. The *AIC* and *BIC* criteria are used to assess model fits, along with the non-nested likelihood ratio test (Ben-Akiva and Lerman, 1985) for evaluating all competing models.

2.7 Variance covariance structure of MHECC Model

The variance-covariance $\mathbf{V}_{I,I}$ formulated from the unobserved heterogeneity for a given segment *q* is a square matrix with dimensions *I*x*I* (*I* = total number or crash types {four in our crash data}), where the variances appear along the diagonal and covariances appear in the off-diagonal elements, as shown below,

$$\mathbf{V}_{I,I} = \begin{bmatrix} \Gamma_1 & & & & \\ \Omega_{(2,1)} & \Gamma_2 & & & \\ \Omega_{(3,1)} & \Omega_{(3,2)} & \Gamma_3 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \Omega_{(I,1)} & \Omega_{(I,2)} & \Omega_{(I,3)} & \Omega_{(I,I-1)} & \Gamma_I \end{bmatrix}. \quad (24)$$

The expected covariance between two independent random continuous variables is estimated directly from the data given by the sum of cross-products formula. This is not the case for the correlated variables where the data are not normally and identically distributed. Hoeffding's formula exists to overcome this difficulty (more details on this formula, see Hoeffding, 1940 and D'Angelo et al., 2013). Hoeffding's formula for the expected covariance between two continuous dependent variables x_i and x_j states that

$$\Omega\left(x_i, x_j\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{C\left[F_i(x_i), F_j\left(x_j\right)|\,\theta_{ij}\right] - \left[F_i(x_i)\times F_j\left(x_j\right)\right]\right\}dx_i dx_j\,. \quad (25)$$

The formula for the expected covariance between two discrete count variables in our case is given as,

$$\begin{aligned}\Omega_{(i,j)}\left(y_{iq}, y_{jq}\right) &= \sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\Pr\left(y_{iq}\le r, y_{jq}\le s\right) - \left[\sum_{r=0}^{\infty}\Pr\left(y_{iq}\le r\right)\right]\times\left[\sum_{s=0}^{\infty}\Pr\left(y_{jq}\le s\right)\right] \\ &= \left\{\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}C\left[F_i(r), F_j(s)|\,\theta_{ijq}\right] - \left[\sum_{r=0}^{\infty}F_i(r)\right]\times\left[\sum_{s=0}^{\infty}F_j(s)\right]\right\}\end{aligned}. \quad (26)$$

More details on Eq. (26) are given in the Appendix A. The average of the expected covariance $\Omega_{(i,j)}$ of all segments is calculated using,

$$E\left[\Omega_{(i,j)}\left(y_i, y_j\right)\right] = \frac{1}{Q}\sum_{q=1}^{Q}\Omega_{(i,j)}\left(y_{iq}, y_{jq}\right) \quad (27)$$

and, the total covariance of crash types *i* and *j* is calculated using,

$$Cov\left(y_i, y_j\right) = Cov\left(\lambda_i, \lambda_j\right) + E\left[\Omega_{(i,j)}\left(y_i, y_j\right)\right]. \quad (28)$$

The variances element Γ_i in the diagonal variance-covariance matrix are calculated for each crash type *i* for the NBII model as, $\Gamma_i\left(y_{iq}\right) = \lambda_{iq} + \psi_{iq}\left(\lambda_{iq}\right)^2$ (the dispersion is given as

$\psi_{iq} = \gamma_i \exp\left(\mathbf{c}_i^T \boldsymbol{\delta}_{iq}\right)$) and $\Gamma_i\left(y_{iq}\right) = \lambda_{iq}$ for the Poisson marginal distributions. The average of the variance Γ_i of all segments is then calculated using,

$$E\left[\Gamma_i\left(y_i\right)\right] = \frac{1}{Q}\sum_{q=1}^{Q}\Gamma_i\left(y_{iq}\right) \tag{29}$$

The total variance magnitude is the sum of two components calculated using,

$$V_T\left(y_T\right) = V\left[E\left(y_T\right)\right] + E\left[V\left(y_T\right)\right] \tag{30}$$

where the $V\left[E\left(y_T\right)\right]$ represents the variance of the expected number of total crash which is constructed from the observed heterogeneity while the second component is the expected variance formulated from the unobserved heterogeneity given in the $\mathbf{V}_{I,I}$ matrix, where both components are given by Eq. (31) and Eq. (32) respectively.

$$V\left[E\left(y_T\right)\right] = \sum_{i=1}^{I} Var\left(\lambda_i\right) + 2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} Cov\left(\lambda_i, \lambda_j\right) \tag{31}$$

$$E\left[V\left(y_T\right)\right] = \sum_{i=1}^{I} E\left[\Gamma_i\left(y_i\right)\right] + 2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} E\left[\Omega_{(i,j)}\left(y_i, y_j\right)\right] \tag{32}$$

3. EMPIRICAL SETTING

The crash data used in the analysis are obtained from Washington State Department of Transportation crash records for three years from 2005 to 2007. Data were collected for Interstate 5 in the state of Washington, USA. In addition to crash data, roadway geometrics and traffic volume data (average daily traffic) were assembled for 274 roadway segments. These segments vary in lengths with roughly 0.87 (miles) mean segment length and 0.60 (miles) standard deviation. These segments include interchange segments only, defined as segments bounded by the farthest ramp terminal on either side of an interchange overpass. The interchanges are spatially separated and regarded as independent from each other, so that the spatial dependency is not considered in this study, but the proposed CML approach can be extended to accommodate it as in Narayanamoorthy et al. (2013). These interchange segments that we considered here, come with many different geometric layouts e.g., directional ramps, semi directional, cloverleaf, diamond, single-point, clove-part, part-diamond and others. For each segment, crashes were recorded by

year and aggregated under each individual type of crash category. Hence, crash frequency counts by types were obtained for each freeway segment for three years resulting in a balanced panel of 822 observations. A total of 13,357 crashes were analyzed in this study. A detailed description of this dataset is provided in Mothafer et al. (2016). While the crash type information included rear end; sideswipe and fixed objects and "all-other," geometric data included: percentage of lanes cross section proportion by length of the segment; central angle of horizontal curves; minimum and maximum radii of horizontal curves; grade; minimum grade; maximum grade; grade differential; number of changes in grade; tangent length; number of horizontal curves per segment; number of vertical curves per segment; presence of exit and entrance.

4. MODEL ESTIMATION AND PERFORMANCE

In this section, we started formulating the empirical copula function for all the crash type pairs. The work was utilized to develop the graphical techniques, PP-plot and the tail-dependence as we explained earlier. Later, we applied the MHOCC model to our dataset. Next, we took the results of both the empirical copula and the MHOCC to develop our last model. The MHECC has confirmed our expectation of detecting the heterogeneity effect of each observed segment, through parametrizing the dispersion and the level of dependency parameters among our different crash types. It is followed by a more investigation on the variance and covariance structure and the correlation among the unobserved heterogeneity that triggered from the joint these crash types. Finally, the marginal effect is also presented to explain the effect of each individual explanatory variable on the crash count by type.

4.1 Empirical copula diagnosis

The empirical copula is formulated using Eq. (19) for each pair of the designated crash types (there are six pairs in total which are: rear-end/sideswipe, rear-end/fixed object, rear-end/ 'all-others', sideswipe/fixed object, sideswipe/ 'all-others' and fixed object/ 'all-others'). The empirical copula[7] of rear end and sideswipe crash types pair is given in Figure (1) (other pairs are not reported in this paper). We used the 1/Q empirical copula for our estimation and it was compared to a selected parametric copula (Gumbel) as an example (other types of empirical copula like

[7] The empirical copula is assumed that only the observations are involved and constructed from a random sample for both the empirical marginals $\tilde{F}_i^n(m_{iq})$, $\tilde{F}_j^n(m_{jq})$ and no explanatory variables are included.

Hazen; Weibull and Bernstein did not differ significantly from our selected empirical copula for all other pairs). We investigated six different types of copula from two different families (elliptical: Independent; Gaussian, Archimedean: Frank; Gumbel; Clayton and Joe) using both the PP-plot and the tail dependence graphs. Figure (2) shows the PP-plot for the rear-end vs sideswipe crash type pair, which we used for the empirical copula and repeated for each parametric copula that we assigned to our developed model later. The figures show that all parametric copulas have plots far from the diagonal line in the smaller probability area (< 0.5) (left and lower side of the graph), but that independent copula has off-diagonal plots also in the larger probability area. The results among all other pairs (other pairs are not reported in this paper) imply that interdependency among crash types is obvious while the empirical copula may not follow any parametric copulas at the segments with no or smaller number of crashes. While PP-plots are useful for comparing cumulative distributions, the weakness of the PP-plots lies in their inability to distinguish important differences concealed by the use of cumulative distributions (Gibbons and Chakraborti, 2011). Therefore, we conducted tests of tail dependence to investigate the tendency of the upper/lower tails of the crash count types distributions. The tail dependence plot depends only on the empirical copula and so it is not restricted to a specific class of copula. The tail dependence of the rear end and the sideswipe crash types is shown in Figure (3). We can see that the empirical copula has a large step around 0.2 at the lower tail, which represents the segments with zero crashes, and smaller fluctuations at the upper tail (segments with high/moderate number of crash count). In the upper tail, most of the observed segments exhibit a pattern closer to Gumbel copula rather than all other types. The same results were observed for other pairs (other pairs are not reported in this paper). Shirazi et al. (2016) stated that the heavy tail crash counts dataset (excessive zero crash count or very large crash count) can cause some problems if the NB regression model is used. We hope that the proposed technique here can clarify the definition of the heavy tail crash counts by separating it into two cases represented by the upper and lower tails as stated.

4.2 Model specification

Let y_{iq} denotes the observed crash count outcome of type *i* and segment *q*, where *i* takes the value of "rear-end" (*i*=1), "sideswipe" (*i*=2), "fixed object" (*i*=3) and "all-others" (*i*=4) respectively. Let also y_{iq}^{*} denotes the unobserved latent tenancies for each crash type correspondingly. We assume

that each crash type follows a NB-II marginal distribution with a specification $F_i(y_{iq})$ and dispersion parameter ψ_i for the MHOCC model. In contrast to the MHOCC mode, the dispersion parameter and the level of dependency parameter are parameterized using Eq. (16) and Eq. (20) when the MHECC model is constructed. Both empirical models consider parameterizing the mean of the expected number of crashes for each type (denoted as λ_{iq}) as a function of all the explanatory variables $\mathbf{x}_{iq}$ with the corresponding parameters $\boldsymbol{\beta}_i$. Furthermore, there are 4(4-1)/2=6 pairs of bivariate probability computations in the CML likelihood function (Eq. (21)), which are: rear-end/sideswipe, rear-end/fixed object, rear-end/ 'all-others', sideswipe/fixed object, sideswipe/ 'all-others' and fixed object/ 'all-others'. To confirm the conclusions that we obtained from the empirical copula, we investigated all types of copula in a sense of goodness-of-fit, through the advantage of our parametric MHOCC model. These parametric copulas are implemented to fit our dataset with unobserved heterogeneity ε_{iq} that generated from each crash type latent variable. The standard normal distribution is selected to represent the continuous variable ε_{iq} and NBII margins are used for the count variable y_{iq} for both the independent and the Gaussian copula.

We begin by identifying the most significant explanatory variables $\mathbf{x}_{iq}$ for each crash type independently due to the fact that each crash type has its distinct mechanisms and characteristics. For this purpose, the MHOCC-independent copula is used where no correlation among crash types is assumed. The independent copula works as a reference to assess both our selection of these explanatory variables and also as a reference to compare when we select different types of parametric copula functions. The preliminary estimation[8] of the MHOCC-independent copula of the independent copula suggests the use of a Poisson marginal for 'all-others' crash type category due to an insignificant dispersion parameter ψ_4 .The estimation results of the MHOCC-independent copula are presented in Table (2). It is worth to mention that the independent copula

[8] Many numerical difficulties arise from the presence of both the gamma function and the factorial function in the negative binomial marginal distribution. These difficulties are realized spontaneously in computing the probabilities if the latter are associated with large crash count number. In GAUSS reference manual (Aptech, 2014), it is stated that maximum allowed number of both gamma/factorial function arguments should not exceed 170. Obviously, this is not the case in our crash count dataset (e.g., in one of the observed interchange segments of the no.5 freeway, rear end crash count is recorded to 212 crashes. Avoiding the overflow problems can be easily achieved if we use the logarithm of both these functions. Thus, to facilitate the speed of computation, Sterling's formula is used, which offers an approximation for these both functions, (in this regard see Winkelmann, (2008) for more details).

can be obtained[9] also by setting all the correlation components θ_{ij} of Gaussian copula to zero.

Gaussian; Frank; Clayton; Gumbel and Joe copulas were evaluated using our MHOCC model. While all of the above parametric copulas indicated significance of correlation with respect to the independent copula, Table (3) provides the log-likelihood, *AIC* and *BIC* measures for each copula model. It is clear that Gumbel copula is the most suitable to fit our data with the highest value of log-likelihood and lower values of *AIC* and *BIC* respectively. The results suggest that the interdependency among crash types is significant as shown in the empirical copula diagnosis. Also, Gumbel copula are better than Frank, Gaussian, Clayton, and Joe copula, implying that the interdependency among crash types are better represented by the asymmetric copula than symmetric counterpart. Conclusively, the result that Gumbel copula is better than all other copulas means the interdependency among crash types has a strong upper tail dependency.

The nested-likelihood ratio test is conducted between the MHOCC-Gaussian and the MHOCC-independent copula, the Gaussian copula collapses to the independent copula by suppressing the dependency parameters among the crash types. The value of the test statistic can be calculated as $109.71\left(=-2\times\left[-17{,}919.91-\left(-17{,}865.06\right)\right]\right)$, which is much greater than the critical value of Chi-square distribution 16.81 at six degrees of freedom for a probability level of 0.999. The test value is statistically significant, which indicates that considering the correlation using Gaussian copula is more preferable rather than the independent copula.

The non-nested likelihood ratio test is also carried out to draw our last conclusion of statistical model selection by comparing the MHOCC-Gumbel copula to the closest competitor MHOCC-Frank copula model (see Ben-Akiva and Lerman, 1985). The difference in the adjusted rho square (also known as McFadden pseudo R square) ($\bar{\rho}_i^2$) value is (0.00202). The probability that this difference between these two competing models is equal to (0.50080). This value is larger than the critical value, which means this difference could have occurred by chance, given as $\Phi\left(-\left[-2\times 2.02E-03\times LL\left(\text{restricted}\right)+\left(38-38\right)\right]^{0.5}\right)$. The critical value of the cumulative probability term, with $LL\left(\text{restricted}\right)=-36{,}179.896$; is equal to $6.48E-34$ which is almost zero,

[9] We can obtain the independent copula density function in a straightforward form by taking the difference of the copula between the two marginals $F(y_1)$ and $F(y_2)$ as $c(F(y_1), F(y_2); \theta) = [F(y_1).\ F(y_2)] - [F(y_1-1).\ F(y_2)] - [F(y_1).\ F(y_2-1)] - [F(y_1-1).\ F(y_2-1)]$,

indicating that the difference in $\bar{\rho}_i^2$ between Gumbel and Frank models is statistically significant to reject the null hypothesis, and that Gumbel copula is more suitable to fit our crash types count data.

We took the results of the work on both the empirical copula and the constant-dispersion parametric copula investigation through the MHOCC model to establish our final model. Gumbel copula is our selection to fit all the crash type pairs; the heterogeneous effect is carried out through parameterizing the dispersion parameters in the MHECC model as we mentioned in section 2.2. MHECC model also includes parameterizing the level of dependency parameter in the copula function as we mentioned in section 2.4. Before developing the MHECC model, we have examined our claim that the heterogeneous dispersion model is comparable to the random parameter models. In this regard, different univariate models of rear end crash type only are presented in Table (4). These models are univariate heterogeneous negative binomial, random parameter Poisson gamma (negative binomial), and random parameter Poisson lognormal. We notice the log-likelihood at convergence of univariate heterogeneous NBII model is closer and related (larger) to the log-likelihood of both random parameters Poisson gamma (negative binomial) and random parameter Poisson lognormal. A second noticeable point arises from comparing the parameter estimates of these three models is that, it seems the random parameter models tend to produce a more significant parameter estimates compared to the heterogeneous ones. This observation can hold for sideswipe, fixed object and 'all-others' crash types which suggests that under the basic level, the heterogeneous model performs similar if not better to the random parameter models. One possible reason one might think of, is that the simulation technique in the random parameter models approximates the probability function, while the heterogeneous approach has a closed-form and no approximation is involved in the estimation.

To make the comparison fair for the same crash type, we have reported in the end of Table (4) the log-likelihood of base models, which are the univariate negative binomial (NBII) (the base of the univariate heterogeneous NBII and random parameters NBII models) and the Poisson lognormal (the base of random parameters Poisson lognormal). We expect that our MHECC model will perform better than the multivariate random parameters for the same reason above. Only the significant explanatory variables in the dispersion function of the univariate heterogeneous NBII model are considered. They are used to parameterize our dispersion functions of each crash type in our multivariate developed model. Thus, no significant variables in the dispersion function of

the sideswipe crash type were found. The dispersion variable of the sideswipe crash type in that case remains a constant across all the given segments.

We parameterized the level of dependency of Gumbel copula in the form $\theta_{ijq}\left(\mathbf{o}_{ijq}\right)=1+\exp\left(\theta_{ij}+\mathbf{d}_{ij}^{T}\,\mathbf{o}_{ijq}\right)$ which is compatible with the range of Gumbel copula $\theta_{ijq}\in[1,\infty)$. The crash type pairwise marginals become independent as θ_{ijq} approaches 1, whereas the pairwise marginals will be strongly correlated and Gumbel copula approaches Fréchet-Hoeffding upper bound as θ_{ijq} goes to infinity (for more details see Bhat et al., 2009). The most significant pairwise variables which influence the dependence level parameter for each crash types pair were selected through several modeling estimations of the MHECC model.

As a final step, and in analogous way to previous, we have compared the two non-nested developed models as an attempt to draw a final conclusion. The difference in the adjusted rho square ($\bar{\rho}_i^2$) value between MHOCC and MHECC model is (0.00182). The probability that this difference between these two competing models is equal to (0.50073). The term $\Phi\left(-\left[-2\times1.82\ E-03\times LL\left(\text{restricted}\right)+\left(57-38\right)\right]^{0.5}\right)$, with $LL\left(\text{restricted}\right)=-36{,}179.896$ is equal to $9.02\ E-31$ which is almost zero, indicating that the difference in $\bar{\rho}_i^2$ between MHOCC and MHECC model is statistically significant to reject the null hypothesis, and that MHECC model, which addresses the heterogeneity, is more suitable to fit our crash types count data.

4.3 Estimation results

Estimation results of MHECC-Gumbel regression model are presented in Table (5) with the variables influencing the dispersion parameters for each crash types. The pairwise variables which influence the dependence level parameter to explain the correlation among the marginal distributions of crash type pairs are presented at the end of the same table and to be discussed later. The estimation results provide parameter estimates for four types of crashes. The dispersion variables across all segments is estimated, with an average 1.649, 5.765 and 5.082 (Min: 0.031, 5.765 and 0.375), (Max: 4.477, 5.765 and 26.055) for rear end, sideswipe and fixed object crash type respectively, which imply an average overdispersion magnitude of 0.606 0.173 and 0.197. The parameters in the dispersion functions are statistically significant for the AADT of both rear end and fixed object as evidence by the large t-values. The MHECC-Gumbel model supports

different sizes of overdispersion for each crash type and over different observed segments. The correlations of the unobserved heterogeneity for each frequency of crash type across all observed segments are also considered in this model.

The MHECC-Gumbel copula indicates unanimously significant relationships between traffic crash type and AADT, segment length and lane cross section proportion (3 lane or greater). The number of horizontal curves variable and the diamond interchange dummy are significant for the rear end, sideswipe and fixed object functions. The positive sign of the horizontal curves parameter indicates that as the number of curves increases, the expected crash count of rear end; sideswipe and fixed object crashes increases as well. The horizontal curve variable captures the effect of speed differentials and lane offsetting on rear end and sideswipe crash likelihoods, and potential loss of control and roadside encroachments resulting in fixed object crash type. The rural indicator variable has a negative coefficient for both the rear end and sideswipe. The diamond interchange indicator reduces the probability of rear end, sideswipe and fixed objects crashes. The minimum vertical grade and maximum vertical curve elevation variables are significant in the sideswipe function.

The horizontal curvature characteristics are represented by the largest horizontal curve central angle in segment (0.515) which is significant at 1% level for the fixed object crash type. The number of vertical curves variable influences the fixed object likelihood in a negative manner, resulting in fewer run off the road crashes involving objects on the roadside.

4.4 Representativeness of variance and covariance structure

As shown in lower part of Table (5), the dependency influential pairwise variables show several significant variables except for the pair of fixed object and other types. For example, ADT is statistically significant for rear-end/sideswipe, rear-end/fixed object, rear-end/all others, sideswipe/fixed object and sideswipe/all others, pairs indicating most likely of these crash pairs to occur. Analogously, the length of the segment is statistically significant for the pairs rear-end/fixed object and sideswipe/fixed object only. Finally, the rural indicator variable, lane cross section proportion and number of horizontal curves are found to be statistically significant for the pairs rear-end/other types and sideswipe/fixed object respectively. As it is pointed by Chandra et al. (2010), it is hard to interpret the influence of the parameter estimate signs of the above variables on the level of dependency function, since the $\mathbf{d}_{ij}^{T}$ vectors incorporated in the exponential function

when we parameterized the Gumbel copula. Each θ_{ijq} dependency parameters represents common unobserved factors in the latent variable functions for each observed segment and for each given crash types pair. Table (6) shows the average, minimum and maximum values of the level of dependency variables, which are estimated using the MHECC model. The average values of θ_{ijq} indicate an association between the unobserved factors of each crash type in the corresponding pair with all pairwise correlations. The non-parametric Kendall's 'τ_{ijq}' measure was utilized to interpret the level of dependency θ_{ijq}, and the results are presented in Table (7). The correlation ranges between 0.075 and 0.203, which demonstrates the presence of common unobserved factors association of the unobserved latent propensity for each crash type. The rear-end, sideswipe and fixed object/other types pairs appear to have weak correlations compared to the other crash type pairs, with the rear-end/sideswipe correlation being the strongest. This is intuitive since they are the same direction, multi-vehicle interactions that occur within lane or in adjacent lanes. One would expect the adjacent and in-lane dynamics to contribute the most to crash type correlations.

The $\mathbf{V}_{I,I}$ matrix was calculated[10] for MHECC-Gumbel copula considering the average values among all segments using both Eq. (27) and Eq. (29) and it's equal to,

$$\mathbf{V}_{4,4} = \begin{bmatrix} 79.78 & & & \\ 47.81 & 5.82 & & \\ 24.21 & 6.32 & 4.01 & \\ 13.73 & 4.14 & 2.31 & 1.66 \end{bmatrix} \tag{33}$$

The total covariance of crash types *i* and *j* is a sum of two components, the covariance resulting from estimated expected number of crash specific type and the one from the association of the stochastic error terms generated from each marginal pair given in Eq. (26). The total covariance of crash types is calculated for MHECC-Gumbel, MHOCC-independent copula models using Eq. (28) and presented in Table (8). The results suggest that MHECC-Gumbel copula represents accurately the covariance structure among the crash types.

[10] It is worth to mention that evaluating the expected covariance elements in Eq. (26) can be also done by Eq. (A-1), but the amount of computation time increases rapidly as the maximum number of crashes for a certain type of crash and a given segment increases. This is because Eq. (A-1) requires to calculate the probability using the differences between the upper and lower bounds for each crash type pair as given in Eq. (21). Theoretically the maximum number of crashes should be set to (+∞) as given by Hoeffding's formula, but we found that a 500 crashes count (upper bound) for each type are adequate to get stability in calculating the expected covariance value for each pair.

The total variance value is also a sum of two components, variance of the expected number of crashes and expected variance among the segments calculated using Eq. (31) and Eq. (32) respectively. Total variance components of MHECC-Gumbel, and MHOCC-independent copula models are presented in Table (9). The results suggest that MHOCC-Gumbel copula also represents the total variance structure more accurately compared to the other model. Finally, we have calculated the estimated correlation for a given segment using the following formula (Ophem, 1999) which are presented in Table (10).

$$\rho\left(y_{iq}, y_{jq}\right) = \frac{\Omega\left(y_{iq}, y_{jq}\right)}{\sigma_{y_{iq}} \times \sigma_{y_{jq}}} \tag{34}$$

4.5 Marginal effects

The marginal effect of an explanatory variable x_v (where $v = 1,2,3....$ represents the number of explanatory variable in the vector **x**) can be obtained by taking the first derivative of the expected number of type specific crash function λ_{iq} with respect to x_v in the MHECC-Gumbel copula model.

$$\frac{\partial \lambda_{iq}}{\partial x_{iv}} = \beta_{iv} \exp\left(\boldsymbol{\beta}_i^T \mathbf{x}_{iq}\right) \tag{35}$$

The marginal effect values of all the explanatory variables of MHECC-Gumbel along with MHOCC-independent copula model for each crash type are presented in Table (11). The marginal effects in the rear end crash type are larger in absolute value than any other crash types regardless of explanatory variables. The results suggest that interstate rear end crash likelihood is more sensitive to geometric and traffic conditions which match the finding of Mothafer et al. (2016). Comparing MHECC-Gumbel copula model and MHOCC-independent copula model, most of the variables have the similar marginal effects within ten percent difference. However, the largest beginning vertical curve elevation in segment for the sideswipe and Number of horizontal curves per segment for fixed object have larger difference in the marginal effect between the two models. These two variables have coefficient estimates not statistically very significant (only at 10% significance level) as shown in Table (11), so the lower accuracy in the parameter estimate might have caused the difference in the estimation of the marginal effects. On the contrary, lane cross section proportion of rear-end, sideswipe and fixed object; largest horizontal curve central angle in segment of fixed object have a very significant coefficient estimate as shown in Table (5), and

the dispersion parameter is also statistically significantly heterogeneous these crash types. Thus, the results suggest that the marginal effect of these variables might be biased if the interdependency among the crash types and the heteroscedasticity of the dispersion parameter are not considered.

5. CONCLUSIONS

This paper presents a multivariate copula-based ordered response model for non-negative integer counts outcomes. The advantages of the proposed model are that it offers a joint distribution without any restrictions on the nature of the correlation (both positive and negative correlations). Second, capability in addressing the variations (heterogeneity) across the observed segments is provided. Third, the need for a simulation-based technique is circumvented. The proposed model uses an alternative way to utilize a latent continuous variable of the ordered response model and match the probability of this latent variable to a corresponding count outcome variable probability. The error term components are assumed as equivalent to the corresponding latent variables that represent different count outcomes. The bivariate copula function in the CML technique is used to pair two count marginal distributions that reflect two different count outcomes. The proposed model is parametric; straightforward to implement and more flexible via allowance for parameterization of the count marginal distribution. The proposed model also offers a non-linear asymmetric interdependence structure among error term components. The correlations among the error components are obtained from transferring the level of dependency of the copula function into a non-parametric Kendall's 'τ' measure.

The above described model framework is demonstrated empirically through the evaluation of dependence among four different crash types that commonly occur on freeway segments located on interstate 5 in the State of Washington. Accounting for the effects of geometry and traffic characteristics of the freeway segments we evaluated five different copula functions using the NB-II marginal. The empirical results show that Gumbel copula is a plausible alternative for accommodating asymmetric tail dependence in heterogeneity among freeway crash types.

APPENDIX A

For discrete count random variables y_{iq} and y_{jq}, any bivariate joint probability cumulative functions C of y_{iq} and y_{jq} with margins F_i and F_j can satisfy the condition,

$$\begin{aligned}\Omega_{(i,j)}\left(y_{iq},y_{jq}\right)&=\left\{E\left(y_{iq}\times y_{jq}\right)-E\left(y_{iq}\right)\times E\left(y_{jq}\right)\right\}\\&=\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\left\{r\times s\times\Pr\left(y_{iq}=r,y_{iq}=s\right)\right\}\\&\quad-\left(\sum_{r=0}^{\infty}r\times\Pr\left(y_{iq}=r\right)\right)\times\left(\sum_{s=0}^{\infty}s\times\Pr\left(y_{jq}=s\right)\right)\end{aligned}\tag{A-1}$$

For any event π of y_{jq}, one has

$$\begin{aligned}\sum_{r=0}^{\infty}r\times\Pr\left(y_{iq}=r,\pi\right)&=\lim_{R\to\infty}\left\{\sum_{r=1}^{R}\Pr\left(y_{iq}=r,\pi\right)+\sum_{r=2}^{R}\Pr\left(y_{iq}=r,\pi\right)+\ldots\right.\\&\quad\left.+\sum_{r=R-1}^{R}\Pr\left(y_{iq}=r,\pi\right)+\Pr\left(y_{iq}=R,\pi\right)\right\}\\&=\lim_{R\to\infty}\left\{\left[\Pr\left(y_{iq}\le R,\pi\right)-\Pr\left(y_{iq}=0,\pi\right)\right]+\left[\Pr\left(y_{iq}\le R,\pi\right)-\Pr\left(y_{iq}\le 1,\pi\right)\right]+\ldots\right.\\&\quad\left.+\left[\Pr\left(y_{iq}\le R,\pi\right)-\Pr\left(y_{iq}\le R-1,\pi\right)\right]\right\}\\&=\lim_{R\to\infty}\left\{R\times\Pr\left(y_{iq}\le R,\pi\right)-\sum_{r=0}^{R-1}\Pr\left(y_{iq}\le r,\pi\right)\right\}\end{aligned}\tag{A-2}$$

similarly,

$$\sum_{s=0}^{\infty}s\times\Pr\left(\mu,y_{jq}=s\right)=\lim_{S\to\infty}\left\{S\times\Pr\left(\mu,y_{jq}\le S\right)-\sum_{s=0}^{S-1}\Pr\left(\mu,y_{jq}\le s\right)\right\}\tag{A-3}$$

for any event μ of y_{iq}. It follows from the identities Eq. (A-2) and Eq. (A-3) that

$$\sum_{r=0}^{\infty}r\times\Pr\left(y_{iq}=r\right)=\lim_{R\to\infty}\left\{R\times\Pr\left(y_{iq}\le R\right)-\sum_{r=0}^{R-1}\Pr\left(y_{iq}\le r\right)\right\}\tag{A-4}$$

$$\sum_{s=0}^{\infty}s\times\Pr\left(y_{jq}=s\right)=\lim_{S\to\infty}\left\{S\times\Pr\left(y_{jq}\le S\right)-\sum_{s=0}^{S-1}\Pr\left(y_{jq}\le s\right)\right\}\tag{A-5}$$

and

$$\begin{aligned}\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} r\times s\times \Pr\left(y_{iq}=r,y_{jq}=s\right) &= \sum_{s=0}^{\infty} s\times\left[\sum_{r=0}^{\infty} r\times \Pr\left(y_{iq}=r,y_{jq}=s\right)\right]\\ &= \lim_{R,S\to\infty}\left\{R\times\sum_{s=0}^{S} s\times\Pr\left(y_{iq}\le R,y_{jq}=s\right)-\sum_{r=0}^{R-1}\sum_{s=0}^{S} s\times\Pr\left(y_{iq}\le r,y_{jq}=s\right)\right\}\\ &= \lim_{R,S\to\infty}\left\{(R\times S)\times\Pr\left(y_{iq}\le R,y_{jq}\le S\right)-R\times\sum_{s=0}^{S-1}\Pr\left(y_{iq}\le R,y_{jq}\le s\right)\right.\\ &\quad \left.-S\times\sum_{r=0}^{R-1}\Pr\left(y_{iq}\le r,y_{jq}\le S\right)+\sum_{r=0}^{R-1}\sum_{s=0}^{S-1}\Pr\left(y_{iq}\le r,y_{jq}\le s\right)\right\}\end{aligned} \tag{A-6}$$

Then we can write the covariance in the form of the copula function as (see Lee, 2001),

$$\begin{aligned}\Omega_{(i,j)}\left(y_{iq},y_{jq}\right) &= \lim_{R,S\to\infty}\left[(R\times S)\times\Pr\left(y_{iq}\le R,y_{jq}\le S\right)-R\times\sum_{s=0}^{S-1}\Pr\left(y_{iq}\le R,y_{jq}\le s\right)\right.\\ &\quad -S\times\sum_{r=0}^{R-1}\Pr\left(y_{iq}\le r,y_{jq}\le S\right)+\sum_{r=0}^{R-1}\sum_{s=0}^{S-1}\Pr\left(y_{iq}\le r,y_{jq}\le s\right)\\ &\quad \left.-\left\{R\times\Pr\left(y_{iq}\le R\right)-\sum_{r=0}^{R-1}\Pr\left(y_{iq}\le r\right)\right\}\left\{S\times\Pr\left(y_{jq}\le S\right)-\sum_{s=0}^{S-1}\Pr\left(y_{jq}\le s\right)\right\}\right]\\ &= \left\{\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} C\left(F_i(r),F_j(s);\theta_{ij}\right)-\left(\sum_{r=0}^{\infty}F_i(r)\right)\times\left(\sum_{s=0}^{\infty}F_j(s)\right)\right\}\end{aligned} \tag{A-7}$$

which is identical to Hoeffding's formula given in Eq. (24) and it can be used to get the covariance between two count dependent random variables.

ACKNOWLEDGEMENTS

This study was partly supported by the joint usage/research program of the Institute of Materials and Systems for Sustainability (IMaSS), Nagoya University. Valuable comments by anonymous reviewers are highly appreciated. Especially, the suggestion for random coefficient modeling by the reviewer at the first-round review resulted in our new proposal for the heterogeneous dispersion modeling.

REFERENCES

Aguero Valverde, J., Jovanis, P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. *Transportation Research Record: Journal of the Transportation Research Board* 2136, 82-91.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident

frequencies with random-parameters count models. *Accident Analysis and Prevention* 41(1), 153-159.

Anastasopoulos, P.C., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash-and non-crash-specific injury data. *Accident Analysis and Prevention* 43(3), 1140-1147.

Anastasopoulos, P.C., Mannering, F.L., 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.

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., Mannering, F.L., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. *Accident Analysis and Prevention* 45, 110-119.

Aptech, 2014. *GAUSS V.14.* Aptech Systems. Maple Valley, Washington.

Asquith, W.H.,2016. *copBasic: General Bivariate Copula Theory and Many Utility Functions. R package version 2.0.4,* Texas Tech University, Lubbock, Texas.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic Methods in Accident Research* 9, 1-15.

Ben-Akiva, M.E., Lerman, S.R., 1985. *Discrete choice analysis: Theory and application to travel demand*. MIT press, 397 pages.

Bhat, C., Astroza, S. and Hamdi, A.S., 2016a. The Formulation and Estimation of a Spatial Skew-Normal Generalized Ordered-Response Model (No. D-STOP/2016/117).

Bhat, C.R. and Dubey, S.K., 2014. A new estimation approach to integrate latent psychological constructs in choice modeling. *Transportation Research Part B: Methodological*, 67, pp.68-85.

Bhat, C.R. and Lavieri, P.S., 2017. A new mixed MNP model accommodating a variety of dependent non-normal coefficient distributions. Theory and Decision, pp.1-37.

Bhat, C.R., 2015. A new generalized heterogeneous data model (GHDM) to jointly model mixed types of dependent variables. *Transportation Research Part B* 79, 50-77.

Bhat, C.R., Astroza, S. and Bhat, A.C., 2016c. On allowing a general form for unobserved heterogeneity in the multiple discrete–continuous probit model: Formulation and application to tourism travel. *Transportation Research Part B: Methodological*, 86,

pp.223-249.

Bhat, C.R., Astroza, S. and Hamdi, A.S., 2017. A spatial generalized ordered-response model with skew normal kernel error terms with an application to bicycling frequency. *Transportation Research Part B: Methodological*, 95, pp.126-148.

Bhat, C.R., Astroza, S., Bhat, A.C., Nagel, K., 2016b. Incorporating a multiple discrete-continuous outcome in the generalized heterogeneous data model: Application to residential self-selection effects analysis in an activity time-use behavior model. *Transportation Research Part B* 91, 52-76.

Bhat, C.R., Astroza, S., Lavieri, P.S., 2017. A new spatial and flexible multivariate random-coefficients model for the analysis of pedestrian injury counts by severity level. *Technical Paper*, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin.

Bhat, C.R., Born, K., Sidharthan, R., Bhat, P.C., 2014a. A count data model with endogenous covariates: formulation and application to roadway crash frequency at intersections. *Analytic Methods in Accident Research* 1, 53-71.

Bhat, C.R., Dubey, S.K. and Nagel, K., 2015. Introducing non-normality of latent psychological constructs in choice modeling with an application to bicyclist route choice. *Transportation Research Part B: Methodological*, 78, pp.341-363.

Bhat, C.R., Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B* 43(7), 749-765.

Bhat, C.R., Paleti, R., Castro, M., 2014b. A new utility-consistent econometric approach to multivariate count data modeling. *Journal of Applied Econometrics* 30(5), 806-825.

Bhat, C.R., Paleti, R., Singh, P., 2014c. A spatial multivariate count model for firm location decisions. *Journal of Regional Science* 54(3), 462-502.

Bhat, C.R., Pinjari, A.R., Dubey, S.K., Hamdi, A.S., 2016b. On accommodating spatial interactions in a generalized heterogeneous data model (GHDM) of mixed types of dependent variables. *Transportation Research Part B* 94, 240-263.

Bhat, C.R., Sener, I.N., 2009. A copula-based closed-form binary logit choice model for accommodating spatial correlation across observational units. *Journal of Geographical Systems* 11(3), 243-272.

Bhat, C.R., Sener, I.N., Eluru, N., 2010. A flexible spatially dependent discrete choice model: formulation and application to teenagers' weekday recreational activity participation. *Transportation Research Part B* 44(8), 903-921.

Boucher, J.P., Denuit, M. and Guillen, M., 2008. Models of insurance claim counts with time dependence based on generalization of Poisson and negative binomial distributions. *Variance* 2(1), 135-162.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., Roncalli, T., 2000. *Copulas for finance-a reading guide and some applications*. Electronically available at http://ssrn.com/abstract=1032533, 69 pages.

Cameron, A.C., Li, T., Trivedi, P.K., Zimmer, D.M., 2004. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal* 7(2), 566-584.

Cameron, A.C., Trivedi, P.K., 2013. *Regression analysis of count data*. Econometric Monograph No.53. Cambridge University Press, 566 pages.

Castro, M., Paleti, R., Bhat, C.R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. *Transportation Research Part B* 46 (1), 253-272.

Castro, M., Paleti, R., Bhat, C.R., 2013. A spatial generalized ordered response model to examine highway crash injury severity. *Accident Analysis and Prevention* 52, 188-203.

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X., Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. *Accident Analysis and Prevention* 99, 330-341.

Chiou, Y.C., Fu, C., 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. *Accident Analysis and Prevention* 50, 73-82.

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.

D'Angelo, G.M., Weissfeld, L.A., 2013. Application of copulas to improve covariance estimation for partial least squares. *Statistics in Medicine* 32(4), 685-696.

Das, A., Abdel-Aty, M.A., 2011. A combined frequency–severity approach for the analysis

of rear-end crashes on urban arterials. *Safety Science* 49(8), 1156-1163.

Deheuvels, P., 1979. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. *Academie Royale de Belgique Bulletin de la Classe des Sciences* 5 *Serie* 65(6), 274-292.

Denuit, M., Lambert, P., 2005. Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis* 93(1), 40-57.

Disanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object–passenger car crashes. *Accident Analysis and Prevention* 34(5), 609-618.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. *Accident Analysis and Prevention* 70, 320-329.

El-Basyouny, K., Sayed, T., 2009. Collision prediction models using multivariate Poisson-lognormal regression. *Accident Analysis and Prevention* 41(4), 820-828.

Ferdous, N., Eluru, N., Bhat, C.R., Meloni, I., 2010. A multivariate ordered-response model system for adults' weekday activity episode generation by activity purpose and social context. *Transportation Research Part B* 44(8), 922-943.

Genest, C., Nešlehová, J., 2007. A primer on copulas for count data. *Astin Bulletin* 37 (2), 475-515.

Gibbons, J.D., Chakraborti, S., 2011. *Nonparametric statistical inference*, Springer Berlin Heidelberg, 977-979.

Godambe, V.P., 1960. An optimum property of regular maximum likelihood estimation. *The Annals of Mathematical Statistics* 31(4), 1208-1211.

Hernández-Maldonado, V., Díaz-Viera, M., Erdely, A., 2012. A joint stochastic simulation method using the Bernstein copula as a flexible tool for modeling nonlinear dependence structures between petrophysical properties. *Journal of Petroleum Science and Engineering* 90, 112-123.

Heydari, S., Fu, L., Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. *Analytic Methods in Accident Research* 13, 16-27.

Hoeffding, W., 1940. Massstabinvariante korrelationstheorie. In *Kommission bei Teubner*,

182-233.

Hüsler, J., Reiss, R.D., 1989. Maxima of normal random vectors: between independence and complete dependence. *Statistics and Probability Letters* 7(4), 283-286.

Imprialou, M.I.M., Quddus, M., Pitfield, D.E., 2015. Predicting the safety impact of a speed limit increase using condition-based multivariate Poisson lognormal regression. *Transportation Planning and Technology*, 1-21.

Joe, H., 1990. Families of min-stable multivariate exponential and multivariate extreme value distributions. *Statistics and Probability Letters* 9(1), 75-81.

Joe, H., 1997. *Multivariate models and multivariate dependence concepts*. CRC Press.

Joe, H., 2014. *Dependence Modeling with Copulas*. Chapman and Hall/CRC Press, 480 pages.

Lavieri, P.S., Bhat, C.R., Pendyala, R.M. and Garikapati, V.M., 2016. Introducing latent psychological constructs in injury severity modeling: multivehicle and multioccupant approach. *Transportation Research Record: Journal of the Transportation Research Board*, (2601), pp.110-118.

Lavieri, P.S., Dias, F.F., Juri, N.R., Kuhr, J. and Bhat, C.R., 2017. A model of ridesourcing demand generation and distribution. Technical paper, Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin.

Lee, E.H., 2014. Copula analysis of correlated counts. In Jeliazkov, I., Poirier, D.J., (ed.) *Bayesian Model Comparison (Advances in Econometrics, Volume 34)*, Emerald Group Publishing Limited, 325-348.

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.

Lee, L.F., 1983. Generalized econometric models with selectivity. *Econometrica: Journal of the Econometric Society* 51(2), 507-512.

Lee, L.F., 2001. On the range of correlation coefficients of bivariate ordered discrete random variables. *Econometric Theory* 17(1), 247-256.

Li, Z., Wang, W., Liu, P., Bai, L., Du, M., 2015. Analysis of Crash Risks by Collision Type at Freeway Diverge Area Using Multivariate Modeling Technique. *ASCE Journal of Transportation Engineering* 141(6), p.04015002.

Lindsay., 1988. Composite likelihood methods. *Contemporary Mathematics* 80, 221-239.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A* 44(5), 291-305.

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.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. *Analytic Methods in Accident Research* 1, 1-22.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research* 11, 1-16.

Marshall, A.W., 1996. Copulas, marginals, and joint distributions. *Institute of Mathematical Statistics Lecture Notes*-Monograph Series, Volume 28, 213-222.

Mothafer, G.I., 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.

Narayanamoorthy S., Paleti R., Bhat C.R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation Research Part B* 55, 245-264.

Nashad, T., Yasmin, S., Eluru, N., Lee, J., Abdel-Aty, M.A., 2016. Joint modeling of pedestrian and bicycle Crashes: Copula-Based Approach. *Transportation Research Record: Journal of the Transportation Research Board* 2601, 119-127.

Nelsen, R.B., 2013. *An Introduction to Copulas*. Springer Series in Statistics, 272 pages.

Nikoloulopoulos, A.K., Karlis, D., 2010. Modeling multivariate count data using copulas. *Communications in Statistics-Simulation and Computation* 39(1), 172-187.

Paleti, R., Bhat, C.R., 2013. The composite marginal likelihood (CML) estimation of panel ordered-response models. *Journal of Choice Modelling* 7, 24-43.

Park, E., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record: Journal of the Transportation Research Board* 2019, 1-6.

Rana, T., Sikder, S., Pinjari, A., 2010. Copula-based method for addressing endogeneity in

models of severity of traffic crash injuries: application to two-vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board* 2147, 75-87.

Sener, I.N., Eluru, N., Bhat, C.R., 2010. On jointly analyzing the physical activity participation levels of individuals in a family unit using a multivariate copula framework. *Journal of Choice Modelling* 3(3), 1-38.

Shi, P., Valdez, E.A., 2014. Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics* 55, 18-29.

Shirazi, M., Lord, D., Dhavala, 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.

Shumway, R.H., Stoffer, D.S., 2011. *Time series analysis and its applications: with R examples*. Third Edition. Springer Series in Statistics, 506 pages.

Sklar, M., 1959. Fonctions de répartition `a n dimensions et leurs marges. *Publications of the Institute of Statistics, University of Paris 8*, 229–231.

Sun, J., Frees, E.W., Rosenberg, M.A., 2008. Heavy-tailed longitudinal data modeling using copulas. *Insurance: Mathematics and Economics* 42(2), pp.817-830.

Tajar, A., Denuit, M., Lambert, P., 2001. Copula-type representation for random couples with Bernoulli margins. *University Catholique de Louvain Institut De Statistique Discussion Paper*, 118.

van Ophem, H., 1999. A general method to estimate correlated discrete random variables. *Econometric Theory* 15, 228-237.

Venkataraman, N., Shankar, V., Blum, J., Hariharan, B., Hong, J., 2016. Transferability Analysis of Heterogeneous Overdispersion Parameter Negative Binomial Crash Models. *Transportation Research Record: Journal of the Transportation Research Board* 2583, 99-109.

Venkataraman, N., Shankar, V., Ulfarsson, G.F., 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.

Wang, K., Ivan, J.N., Ravishanker, N., Jackson, E., 2017. Multivariate Poisson lognormal modeling of crashes by type and severity on rural two lane highways. *Accident Analysis and Prevention* 99, 6-19.

Winkelmann, R., 2008. *Econometric Analysis of Count Data*. Fifth Edition, Springer Science and Business Media, 320 pages.

Winkelmann, R., 2012. Copula bivariate probit models: with an application to medical expenditures. *Health Economics* 21(12), 1444-1455.

Yamamoto, T., Morikawa, T., 2013. Development of shopping frequency model considering competition among commercial areas: Application to analysis on changes in shopping behavior after department store opening at city center. (In Japanese). *Journal of the City Planning Institute of Japan* 48(3), 459–464.

Ye, X., Pendyala, R.M., Shankar, V., Konduri, K.C., 2013. A simultaneous equations model of crash frequency by severity level for freeway sections. *Accident Analysis and Prevention* 57, 140-149.

Zeng, Q., Wen, H., Huang, H., Pei, X., Wong, S.C., 2017. A multivariate random-parameters Tobit model for analyzing highway crash rates by injury severity. *Accident Analysis and Prevention* 99, 184-191.

Zhan, X., Aziz, H.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.

Zimmer, D.M., Trivedi, P.K., 2006. Using trivariate copulas to model sample selection and treatment effects: application to family health care demand. *Journal of Business and Economic Statistics* 24(1), 63-76.