Analysis on Crash Types Frequency Models Considering Correlation

Mothafer Ghasak I. M. A.

Analysis on Crash Types Frequency Models Considering Correlations

Doctoral Dissertation

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Engineering

by:
Mothafer, Ghasak Ibrahim Mohamed Amen

Academic Adviser:
Professor Yamamoto, Toshiyuki

Department of Civil Engineering
Nagoya University
JAPAN
December, 2016

Acknowledgments

I would like to express my sincere gratitude to everyone who have contributed and extended support during the completion of this thesis. Firstly and Foremost, I would like to express the deepest appreciation and thanks to my major advisor, Professor Yamamoto Toshiyuki, for his continued and untiring support towards my Ph.D. research. His excellent guidance, patience, enthusiasm, immense knowledge, and faith in me throughout this process have been extremely helpful to finish this work. Prof. Yamamoto is a truly an inspiration for me. Secondly I would like to thank our laboratory leader Professor Morikawa Takayuki, he has always been considered as a great asset in our laboratory, his charisma, knowledge, generous support and advice have been always immensely available to all of us equally, I can't thank him enough for his patience and support during my work, and to his continuous endeavor to make me something in the academic world. Thirdly, I would like to thank Prof. Miwa, Tomio, without his tremendous effort I wouldn't achieve the work here, since he was one of the reasons of me being accepted in NUTREND laboratory, I can't find enough words to thank him, without his support on this issue I almost given it up and back to my country empty handed. I would like to thank associate Professor Kai Liu; associate professor Kato, Hirokazu for serving as my reviewed committee members. I want to thank you all for letting my defense be an enjoyable moment as well as for your brilliant comments and suggestions.

My gratitude further extends to lecturer Dr.-Eng, Sato, Hitomi, for her great support during my graduate work, moreover, my deep thanks extend to all members of Morikawa Yamamoto and Miwa lab (NUTREND) laboratory, including my colleagues and friends, especially those who have been assenting, helping and sharing their joys and distresses with me during my staying in Nagoya University.

Next on my list to thank is the Japanese Ministry of Education, Culture, Sports, Science and Technology for providing me the scholarship, which helped me to focus on my study. Moreover, I would like to express my gratitude to the staffs of Department of Civil Engineering, Nagoya University, especially Mrs. Kawahara, Hiroko.

I really stand today on a solid ground of confidence because of all the

support that I obtained from my beloved family, I can’t express my feelings of appreciation to thank enough my mother and father, for all the sacrifices, endless love and support throughout my life, especially they were the source of my inspiration as I hurdle all the obstacles in the completion of this graduate work. I would like to thank also my beloved brothers: Shafak, Rizq and Amen and my lovely sister Sarah, I am so lucky to have you all on myside.

To my beloved wife Ito, Eriko, I would like to express my deepest appreciation and gratitude to your terrific, marvelous support. You are the one who makes me achieve my dream and finishing my PhD. Your words, compassion, love and encouragement were the energy that I needed to keep it forward. You have explained and still to me both, the beauty of the Japanese intricate philosophy and the sophisticated culture of Japan, this knowledge was so important for me to shape me of who am I today. I would like to thank also my mother-in-law, without her continuous support and help I wouldn’t achieve this work.

Last but not least, I am indebted to my friend Al Nuaimi Ahmed. His remarkable insights and support were more than adequate to solve many problems I faced during modeling and programming. Also I would like to thank all my friends outside of the university, especially, Soma san, Yoshika family, and Watanabe san, you are the resort that I was relying on when I tried to relax myself and to destress myself from the piled works, to all of you, thank you very much.

Abstract

Despite the evident unobserved heterogeneity correlation among the crash types that frequently occur on freeway segments, inadequate research has been devoted in safety analysis to accommodate such correlation. Furthermore, ignoring such correlation could possibly lead to an enormous misleading conclusions and judgments since the former affects the model parameter efficiency. The correlation components in dynamic states alter with the length of the observation time which makes even more difficult to trace. Modeling the unobserved heterogeneity can improve the predictions of the count outcomes of interest as well. Thus, these improvements can be achieved via introducing the multivariate count model concept. Recent advancements in multivariate count econometric models have allowed researchers to investigate the correlations using simulation based techniques which are not so tractable in a sense of time consuming and efficiency. Our main objective in this research is to find more flexible model and easy to be used by analysts, then apply the obtained knowledge to model the traffic crash types counts that frequently occur on freeway segments. We will investigate the correlations and covariances among the rear end, sideswipe, fixed object and other crash types on freeway sections using three-year crash data for 274 multilane freeway segments in the State of Washington, U.S.A.

To comprehend correlations among different types of accidents and explanatory variables, while taking full benefit of the available crash count record, a multivariate Poisson gamma mixture (MVPGM) count model has been implemented. The model consents a restricted correlation pattern allowing for positive correlation among accident types. The model parameters are estimated using a maximum likelihood approach. Based on the empirical results presented in this study error correlations across accidents types are significantly presented. These significant error correlations occur due to common unobserved heterogeneity affecting the specific accident frequency type. The proposed model shows significant unobserved correlations among different types of accidents frequencies. It also provides a better representativeness for the variance and covariance structure of each accident

type Furthermore, the results reveal that rear end accident type is more likely to be affected by geometric and traffic characteristics of freeway. It is also found that considering the types of accidents is efficient and similar to modeling total number of accidents.

For the second part of this study we tried to contribute to methodology enhancement of the multivariate count data modeling by introducing a simple and practical formula. The formulation started from modifying 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 using the simulation maximum-likelihood to solve the joint probability of multi-integrations among the count outcomes dimensions in the joint probability function, we proposed to utilize the composite marginal likelihood (CML) approach. It is proven that this approach with the copula formulation works efficiently and easy to be implemented for the discrete data. The proposed model allows the positive and negative dependency among the count outcomes as well as a variety of dependent structures including radially asymmetric or tail dependency without a need for a simulation mechanism.

We apply these techniques to study the interdependence structure for the same crash count dataset. The developed second model is applicable for parameter estimates using the maximum likelihood approach. The empirical results show a significant presence of the unobserved heterogeneity dependency across these types of crashes. The results also show that considering the unobserved heterogeneity are highly recommended to enhance the covariance and the variance structure estimation when they are compared to the observed ones. Another finding is that the characteristics of the horizontal curves on the designated freeway segment increase the likelihood of these types of crashes occurring, when compared to the characteristics of vertical curves.

Later, we shifted our scope to the serial correlation problem using the same crash-count data set that we used before but this time considering the time of observation. The unobserved heterogeneity now is in dynamic status, thus, time invariant heterogeneity arising through multiple years of observation

(between 2005 and 2007) for each segment is viewed as a common unobserved effect at the segment level, and typically treated with panel models involving fixed or random effects. Random effects model unobserved heterogeneity through the error term, typically following a gamma or normal distributions. We exploited the fact that gamma heterogeneity in a multi-period Poisson count modeling framework is equivalent to a negative binomial distribution for a dependent variable which is the summation of crashes across years. The Poisson panel model is the random effects Poisson gamma (REPG). In the REPG model, the dependent variable is an annual count of crashes of a specific type. The multi-year crash sum model is a negative binomial (NB) model that is based on three consecutive years of crash data (2005-2007). In the multi-year crash sum model, the dependent variable is the sum of crashes of a specific type for the three-year period. Four categories (in addition to total crashes) of crash types are considered in this study including rear end, sideswipe, fixed objects and all-other types. The empirical results show that the three-year crash sum model is a computationally simpler alternative to a panel model for modeling time invariant heterogeneity while imposing fewer data requirements such as annual measurements.

Within panel cash-count context, as our final target in this thesis, we utilized all the knowledge we gain through all the developed previous models to construct an econometric framework to model the multivariate panel crash count by type data. The point of emphasis is that modeling multivariate count panel data has more superior econometric benefits, which is clarified in producing more efficient parameter estimates compared to the ones arising from the multivariate cross-sectional models. Therefore, we considered the intertemporal (serial) correlations of a given crash type among the years of observations. Moreover, we have considered the inter-type correlations formulated by jointing the probabilities among different crash types. Both of these correlation components added a higher intricacy to seek a conceivable inference. We developed two flexible models to overcome this problem: Multivariate Panel Poisson Gamma Copula (MVPPGC) and Multivariate Panel Copula-Copula (MVPCC) model. These two models are in no need for a simulation mechanism, which is a common issue to model the multivariate

count outcomes. The source of flexibility of these models is demonstrated through allowing a non-linear asymmetric shape of these correlation components generated among the unobserved heterogeneity of each crash type and across the years of observations. The empirical results suggest that Frank copula statistically outperforms other copula types to fit the serial correlation among the years of observations of each crash type. Moreover, MVPCC model offers a better prediction of the crash-type count, since it more accurately represents the variance-covariance structure.

Table of content

List of Figures

List of Tables

Chapter 1

Introduction

1.1 BACKGROUND

One of the most imperative duties of highway safety consultants is the identification of locations in need of engineering improvements to reduce the number of crashes on a traffic facility. The definition of the traffic crash consists a collisions between two vehicles or vehicle with pedestrian, animal, road debris or other stationary obstruction. A distinctive study by Rumar (1985), using British and American crash reports as data, found that 57% of crashes were due solely to driver factors, 27% to combined roadway and driver factors, 6% to combined vehicle and driver factors, 3% solely to roadway factors, 3% to combined roadway, driver, and vehicle factors, 2% solely to vehicle factors, and 1% to combined roadway and vehicle factors (see Lum and Reagan, 1995).Our motive in this study has been embraced through an attempt to answer the following questions:

- Are we being able in our current of crash models to capture enough the traffic crash phenomenon intricacy?
- Which element of the data that we are unable to represent in our model?
- How to obtain greater insight into traffic crashes and their causes.
- Which road entity (segment) can be called ‘safe’? And why?
- Which part of crashes we are being able to reduce through engineering versus the ones reduced through behavioral investments?

The fundamental key is to connect based on a robust statistical model

some explanatory variables that can reflect (the real traffic; geometric; weather; driver characteristics) conditions to the crash occurrence in order to provide better guidance for policies and countermeasures that will help at reducing the number of crashes. In order to do that we need to thoroughly understand the mechanism of the crash and the correlation among different types of crashes that dominate the high percentage of crash numbers.

There have been a number of papers in the literature to cast the light on the crash types (see Mannering and Bhat, 2014 and Lord and Mannering, 2010). Such types include, rear-end, sideswipe, fixed objects, same direction, opposite direction, head-on, and many other types. Each crash type is uniquely different in its nature and mechanism that distinguishes it from other types. Many of these studies efforts that originally connect crash types with factors (explanatory variables) that reflect the increasing likelihood of crash occurrence have since been already developed. A broad-spectrum of these factors are mainly classified into the following groups: a) human b) vehicle c) geometric d) weather and e) traffic characteristics. A successful econometric model is the one that can predict the number of a specific crash while reflecting these variables in a systematic scheme. Kim et al. (2006) have stated that there are at least three imperative thoughts to develop a crash model by type as a function of these explanatory variables: The first reason is to investigate sites that are considered a high risk for a specific crash type, such information is usually concealed if the total crash count number is used. A second reason is to gain a better understanding on performing suitable countermeasures through knowing the differences in the effects of geometric, traffic, and environmental factors on crash type. Finally, and related to the second, a prior knowledge that we can draw a comparison among these different types. For the all mentioned reasons above, the crash type estimation models provoke intuitions and clear ambiguity regarding crash occurrence, with an insights on providing necessary counteract remedies. Ye *et al.* (2013) modeled crash frequency by severity at freeway using a simultaneous equations Poisson-lognormal model with error components that are normally distributed. Chiou and Fu (2013) implemented multinomial-generalized Poisson model with error components. In both of these two studies, the maximum simulated likelihood

estimation approach has been used. Park and Lord (2007); El-Basyouny and Sayed (2009); Lee et al. (2014); Lee et al. (2015); Li et al. (2015); Ma et al. (2006); Ma et al. (2008); Aguero et al. (2009) and Imprialou et al. (2016) have utilized the Bayesian approach instead. Regardless which approach is used, there is some doubt whether these models could be computationally tractable and less time consumption to obtain solution when applied to high-dimensional multivariate data (Winkelmann, 2008). Other studies considered crash types as an explanatory variable (see Chiou and Fu, 2013; Shaheed *et al.*, 2013; Gkritza *et al.*, 2010; Yan *et al.*, 2011; Yang *et al.*, 2011; Shankar and Mannering, 1996), while other comprehensively studied one type of crashes at a specific facility (Das and Abdel-Aty, 2011; Dissanayake and Lu, 2002).

Count data models stand on the base of a discrete probability distribution theory, the mean of the discrete distribution is parameterized as a function of explanatory variables (Castro et.al, 2012). Evolution of frequency and severity of crashes modeling is still developing. In particular, the effect of geometrics on the crashes. The problem of loss of “efficiency” of parameters when correlation among unobserved factors are considered as associated burden to the empirical models development (Mannering and Bhat, 2014). Such factors can mainly classified into three sources based on where they are triggered from. These sources are: A) Driver factors (such as age, gender, marital status, socioeconomic status, risk taking, driving experience, driving behavior, driving adjustment in situational responses). B) Vehicle factors (such vehicle type, engine type, safety features (airbags, anti-lock, brakes, etc)). C) Road factors (such as local pavement condition, local distractions (billboards, glare, signs…etc)). Engineering efforts lie on including these factors in a form of stochastic error term and include it into a statistical model.

1.2 PROBLEM STATEMENT

On a specific freeway segment several accident types occur due to various factors. Some of these types are likely to be correlated to each other. The need for modeling crash frequencies by accident types has been distinguished. Development of such model to consider these correlations systems are associated with intricacy in formulation/computational aspects. Since crash related outcomes like the Rear-end crash type, Sideswipe or overturn type

exhibit interdependencies (correlations), their simultaneous analysis requires that these interdependencies to be taken into consideration. In view of the fact that these interdependencies among different crash types are related to how to formulate the variance-covariance structure of the unobserved factors, a more investigation on that field need to be carried out.

1.3 OBJECTIVES

The purpose of this study is to develop an empirical statistical system of models that can forecast the frequency of crashes on a specified free highway segments. Then to assess the factors that maximize the likelihood of accident occurrence. Furthermore, to investigate the relationship among crash types in order to get better understand of the latent factors that correlated them together. In more details, we can summarize our objectives into:

1. Demonstrate (as theorized) there are several ways of associated exogenous variables to interact in different way with the designated crash types.
2. Investigate the source of associated problems arise from unobserved heterogeneity (such as overdispersion and the serial correlation (intertemporal effect) problems).
3. Investigate on whether the unobserved heterogeneities are correlated among crash types and getting more efficient predictors.
4. Seeking for a better representation of the variance-covariance structure in order to get better predication.

1.4 RESEARCH OUTLINE

This dissertation comprises an eight chapters. Here we briefly presents the outline of this dissertation which consists our main focus in each chapter. We will give the references among all these chapters under a separated section called 'bibliography' in the end.

In chapter two we will present the main methodological/data associated issues which appear in modeling the crash-count data. These issues are explained shortly into different sections. After that, crash-count available models were discussed to understand the state-of-the-art of the available

developed models in the literature, with their advantages and disadvantages. This point will accompany each chapter in a separate section under the name 'background' which extensively discuss these models and their features. The merits/types of the copula function (mainly the bivariate ones) is explained after, given a reasonable justification of its usage in our developed crash-count models context.

Chapter three presents the our crash-count data, including the site, basic counts tables, crash count type selection, their empirical distributions and a descriptive static of the observed crashes with their associated explanatory variables.

Our crash-count modeling efforts started from chapter four and continued until chapter seven. Each chapter represents an independent work but complements each other in a way of seeking better model performance and efficient parameter estimates. In precise order, chapter four and chapter five deal with cross-sectional count data, while chapter six and chapter seven deal with panel count data. All the developed models are multivariate models type, thus accordingly.

Chapter four, the multivariate Poisson gamma mixture model is developed. This work has yielded as our first journal paper which the crash type covariances and roadway geometric marginals effects have been investigated thoroughly. This model was extended later and presented in chapter five were the bivariate copula function is incorporated in the CML technique. The model shows many beneficial features in modeling our crash count data. In the next chapter, namely chapter six, we have introduced the time of observation effect for first time, thus the random effect Poisson model is used for this purpose. The results were compared to a cross sectional crash sum model and both the advantages and disadvantages between these two models are discussed. We finished our modeling efforts with chapter seven, many concepts from chapter four five and six are incorporated in this chapter. The presented framework is used later to explain the correlation structure among both the crash types and the time of observations. Chapter eight gives a brief conclusions and future work of each chapter in this study (see Figure (1-1).

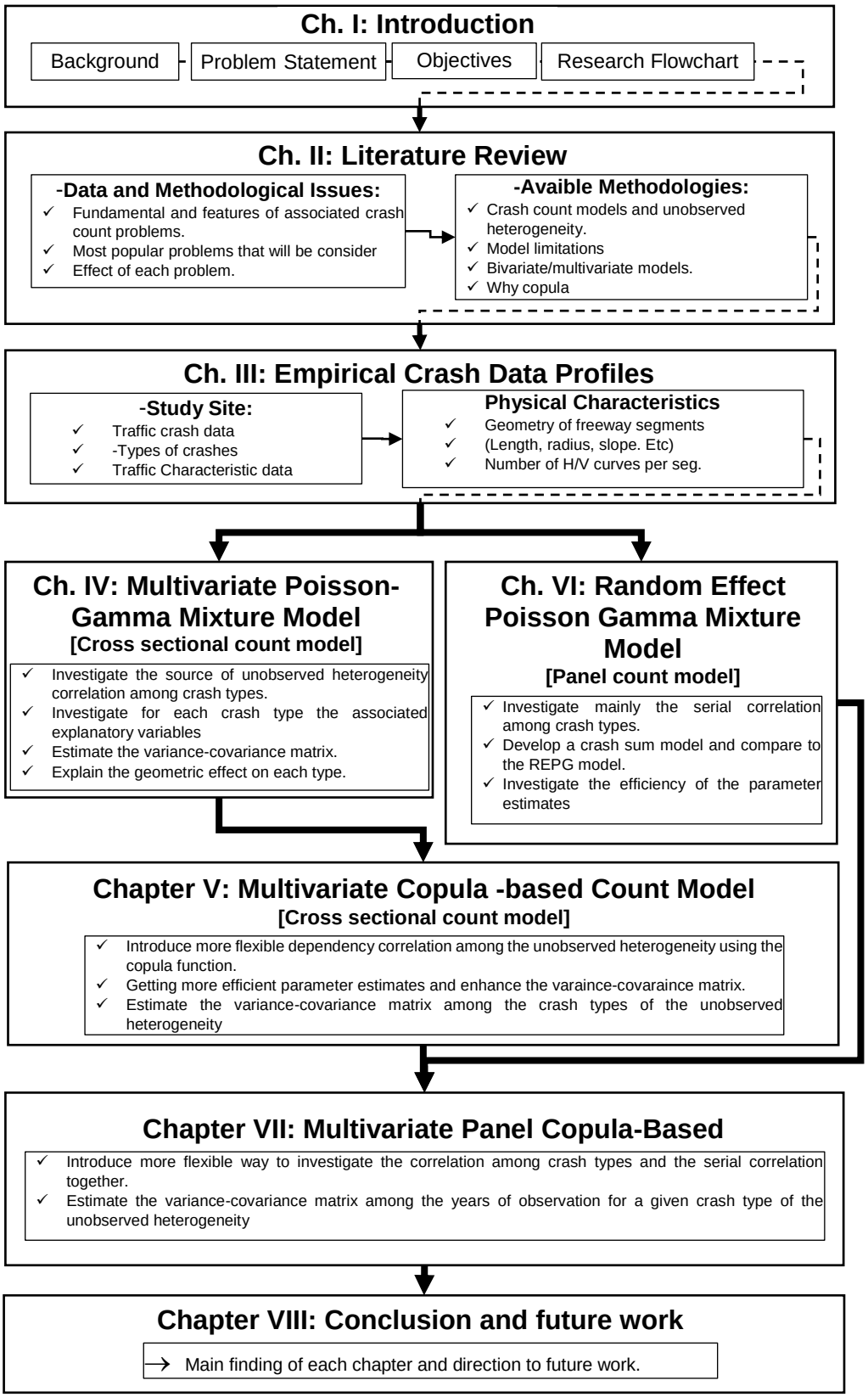


Figure (1-1) Research flowchart

Chapter 2

Literature Review

2.1 INTRODUCTION

Crash count modeling is broadly well-known in the literature see Lord and Mannering, (2010). Most of developed crash count models are used to cover wide range of crashes data and methodological issues. This chapter will first navigate elaborately and shortly all the current data and methodological issues in crash count context, then moves to explore all related developed crash count models that appear in the recent literature.

2.2 DATA AND METHODOLOGICAL ISSUES

Here, we will discuss the major associated problems with crash-frequency (count) that have been recognized over the years in the literature. These problems are considered as a crucial source of error that usually leads to select incorrect statistical model. Thus, this error affects crash-frequency prediction and interpretation of the associated parameter estimates. We will explore these actuarial issues first, then later we will extensively explain the problems in their context in this study. These problems are can be classified as (see for more details (Lord and Mannering 2010) and Mannering, Shankar et al. (2016)):

2.2.1 Over-dispersion

This common problem occurs when the variance value excesses the mean value of the crash counts data. In that case overdispersion will cause violation the equality assumption between these two values which is one of most important assumptions in the existed approaches of count-data modeling (especially if Poisson regression model is used). The estimation of the parameters in that regard will be bias and misleading the inference on the

significant level. The Over-dispersion is generated when there is an unobserved heterogeneity (unobserved factors) that affects the crash-count outcome. Over-dispersion usually a term used in the cross-sectional data reference (which we will discuss later).

2.2.2 Under-dispersion

This problem is not so common in crash-count data especially when the sample size is low. This problem occur when the variance value less than the mean of the crash-counts on the designated roadway segment.

2.2.3 Dynamic explanatory variables (changing with time)

Collecting crash-count data usually involve counting process over specific time period. Most of the explanatory (exogenous) variables are considered a time invariant variables over time. The lack of information on how these variables changing with time is still dominating the crash-count data. In our study we have only the traffic dynamic (represented by the variable (AADT)) that is considered as time-varying variable. This problem are also known in the literature as a serial correlation, autoregressive, intertemporal correlation effect problem.

2.2.4 Temporal and spatial correlation

When a roadway individual entities (segments) is observed over short time periods to evade from lacking of information on the explanatory variables changing with time (the 2.2.3 problem), there is high possibility that these observations are statistically correlated. The source of correlation is generated from the associations of same unobserved heterogeneity of same roadway individual over time/space. This problem is in similar to the over-dispersion problem which leads usually to violate the equality property in Poisson model that we mentioned before.

2.2.5 Small sample size and low sample-mean.

Collecting crash-count data is associated with large cost/time in general. Therefore, these data are collected over a short time span usually. It's an evident that if the sample size small, it will affect highly the safety judgment in general. For example if an analyst prefers to use the zero-inflated count model on an individual roadway with high risk property, it's inevitable to reach to a false judgment. Same goes for small sample size.

2.2.6 Crash-type and injury-severity correlation

If the crash-count data classified into different categories in injury severity level or different crash-type and then modeled in separate forms, it's highly possible to get inefficient parameter estimates of the explanatory variables due to neglecting the correlation among these categories.

2.2.7 Under-reporting

Under report problem appears in lack of recording less sever crashes which leads to an estimate bias usually. For each severity level, rate of underreporting crashes is still unknown in general.

2.2.8 Omitted-variables bias

Developing a model with less explanatory variables in the crash-count estimation function is persuasive (for simplicity and lack of information purposes). Therefore, omitting variables is a common problem in crash-count modeling which leads usually to get biased parameter estimates. Omitting variables causes both the over-dispersion/serial correlation problems.

2.2.9 Endogenous variables

Endogeneity occurs when one of the explanatory or more is related to another explanatory variable in the functional form. Ignoring the endogeneity results a parameter estimate biased usually. Accounting the endogeneity problem in crash-count data adds more complexity in molding

2.2.10 Functional form

Establishing a crash-count model is involved an extra cautious to represent the number of crashes as a cardinal positive, integer number. A non-linear function is preferable in that aspect which adds more complexity especially it may require numerical solutions. Misspecification of the functional form can lead also to over-dispersion/serial correlation problems.

2.2.11 Fixed parameters

In the traditional crash-count modeling, the explanatory variables are treated as fixed-status variables across all the roadway individual entities. In some cases this will lead to erroneous inference unless a dynamic status is introduced which is adds more complexity in molding.

2.3 CRASH-COUNT AVAILABLE MODELS

In context of previous prevailing issues in crash count data, there have been many models that accommodate these problems with advantage/disadvantage that can be found extensively in Lord and Mannering (2010). Various statistical models have been developed that identify factor contributing to crash frequency or severity. The crash count models in the literature can be classified into two classes based on the number of crash count outcomes that can be handled when construct the model. These classes are: a) univariate crash count models and b) multivariate crash count models. It worth to mention that both of these classes can be classified into two classes, which are based on how we are looking into the data. These two classes are cross-sectional crash count models and panel crash count models. We will list here all the available crash count models that dominate the literature.

2.3.1 Poisson model

This model is the most basic of crash count data analysis. Most desirable feature of this model is: it is easy to estimate. The disadvantage of this model lies in lacking of representing the over/under dispersion of the unobserved heterogeneity, it's also affected by the low sample mean and low sample size of the data e.g., Jovanis and Chang (1986), Joshua and Garber (1990), Jones et al. (1991), Miaou and Lum (1993), and Miaou (1994).

2.3.2 Negative binomial (Poisson gamma mixture model)

Negative binomial model is the second basic model and easy to estimate, offers a closed-form of probability density function and can address the overdispersion problem. This model cannot handle the under dispersion problem also it's also affected by the low sample mean and low sample size of the data e.g., Maycock and Hall (1984), Hauer et al. (1988), Brüde and Larsson (1993), Bonneson and McCoy (1993), Miaou (1994), Persaud (1994), Kumala (1995), Shankar et al. (1995).

2.3.3 Poisson-lognormal Model

This model in similar shape to the negative binomial model but rather than using the gamma distribution of the unobserved heterogeneity in the model structure, the lognormal distribution is used instead. Since this model doesn't

offer a closed-form of the probability density function, this model needs a creative technique to obtain solution. In similar to previous models, it's also affected by the low sample mean and low sample size of the data e.g., Miaou et al. (2005), Lord and Miranda-Moreno (2008), and Aguero-Valverde and Jovanis (1989).

2.3.4 Zero-Inflated Poisson and Negative binomial models

This model is used to compensate the lacking of representing zero count in previous models. The zero-inflated model should be used with caution which is necessary not to reach to a wrong conclusion, especially zero count interpretation can reflect the specified roadway entity (segment in our data) how safe it is. The proper observation time unit is vital in that case to give valid conclusion on safety level of the specified entity e.g., Miaou (1994), Shankar et al. (1997), Carson and Mannering (2001), Lee and Mannering (2002), Kumara and Chin (2003), Shankar et al. (1997).

2.3.5 Conway-Maxwell-Poisson model

This model is used to overcome both over and under dispersion problem in the crash count data. In similar to previous models, it's also affected by the low sample mean and low sample size of the data e.g., Lord et al. (2008) and Lord et al. (2010).

2.3.6 Gamma model

This model can handle both over and under dispersion problem in the crash count data. This model has a limited use since it is constructed in duel-state of zero crash count and non-zero e.g., Oh et al. (2006) and Daniels et al. (2010).

2.3.7 Generalized estimating equation model

This model is used to represent the temporal correlation in the panel crash count data context. This model may need an extra information on specifying the types of the temporal correlation e.g., Lord and Persaud (2000), Lord et al. (2005), Halekoh et al. (2006).

2.3.8 Generalized additive models

These models are more flexible to address the non-linearity of variable interactions. The complexity and not easy to transfer to other datasets are considered the main disadvantages of utilizing such models e.g., Xie and

Zhang (2008) and Li et al. (2009).

2.3.9 Random-effects models

These models can be used to represent both the spatial and temporal effect in the dataset. It may not be easy to transfer to other datasets e.g., Johansson (1996), Shankar et al. (1998), Miaou and Lord (2003).

2.3.10 Negative multinomial model

This model is in similar to the Poisson gamma mixture model which is used to investigate the temporal effect (the serial correlation problem). The overdispersion in the data is also represented in in this model which is caused by the temporal effect. This model cannot handle the under dispersion problem also it's also affected by the low sample mean and low sample size of the data e.g., Ulfarsson and Shankar (2003), Shankar et al, 1998 and Sittikariya et al. (2005).

2.3.11 Random-parameters models

These models are more flexible and offer distributions to each explanatory variable rather than fixed values. These models are more complex to estimate and require more creative techniques to get solution since no closed-form available usually e.g., Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009).

2.3.12 Bivariate and multivariate models.

These models are used to model more than one crash outcome (crash types in this thesis) simultaneously. Most distinctive feature of formulating these model is that we need to construct the correlation matrix among these crash count outcomes. The formulated joint probability is in no closed-form, thus a creative technique is needed to obtain solution. These models also more complex to estimate rather the univariate models, but they often offer more insights on how these crash count outcomes interact to each other e.g., Miaou and Lord (2003), Miaou and Song (2005), N'Guessan and Langrand (2005a), N'Guessan and Langrand (2005b), Bijleveld (2005), Song et al. (2006).

2.3.13 Finite Mixture and Markov switching models

These models are used to analyze the source of dispersion in the crash count data but not easy to estimate and not be transferable to other datasets e.g.,

Malyshkina et al. (2009), Park and Lord (2009).

2.3.14 Duration models

These models consider the time duration among crash occurring which offer more analysis depth, but require more details in the crash count data, such as the time-varying of the explanatory variables which is more difficult and expensive to obtain usually e.g., Jovanis and Chang (1989), Chang and Jovanis (1990).

2.3.15 Hierarchical and multilevel models

Classing the crash count outcomes into groups offered by these models give more capabilities to address temporal, spatial and other type of correlations among these groups, but often the correlation is not easy to interpret and not easy to transfer to other datasets e.g., Jones and Jørgensen (2003) and Kim et al. (2007).

2.3.16 Neural, Bayesian neural network, support vector machine models

These models are non-parametric and in no need to any assumption on the data distributions, more flexible to achieve perfect fitting usually but complex to estimate, not easy to interpret the results and may not be transferable to other datasets e.g., Abdelwahab and Abdel-Aty (2002), Chang (2005).

2.4 KEY FEATURES OF OUR APPROACH

Most of these studies mentioned above have neglected the potential source of correlation that may exist across different crash types that will cause “loss of efficiency, thus our contribution in this study will be mainly directed toward this point. Since we considered crash-type correlation and time temporal problems, which both are involved with multivariate-base modeling, we will focus on construct in shape of bivariate/multivariate models. These models are necessary in crash-count modeling when, the joint-probability among different crash count outcomes are the point of interest rather than the ones from total number of crashes. Basically, the correlation among the unobserved factors of each individual crash-count outcome plays a vital role on how to construct these models. As we stated before, Multivariate models are more valuable since they offer more information on how different crash-count outcomes correlated. Multivariate models are better in representing the error term

structure which leads to a better predication usually. Multivariate models are more complex to be estimated especially when several crash-count outcomes are involved in estimation (larger than 6, see Bhat et al. (2014b)). From the analysis complexity and the computational challenges of the multivariate models we concern on offering a better solution since most of the current models are: a) too restricted and allowing only positive correlation b) set to be a linear-symmetric correlation usually. c) Most of the available solutions are computationally extensive, thus the maximum simulated likelihood is often used to obtain solution. Finally, d) the multivariate panel count model is still considered as a big challenge that is not so much considered regardless to its importance in the literature. We will discuss thoroughly in Chapter 4, Chapter 5, Chapter 6 and Chapter 7 all these available models in their context with their advantages and disadvantages.

Chapter 3

Empirical Crash Data Profiles

3.1 CRASH DATA SOURCE

The crash dataset is obtained for interstate 5 in the State of Washington, USA (see Figure (3-1)). The interstate no.5 is a multi-lane divided highway that established in august 7th, 1947 by FHWA and entered in operational service from 1956 until today. It is the only interstate highway to traverse the whole north-south length of State of Washington, serving the major cities like Vancouver Olympia, Tacoma and Seattle.

Three years of crash data were collected from 2005 to 2007. Data contained three different categories: (1) the crash count of the total number of crashes and each type of crashes data; (2) the geometrical characteristics of these highways; and (3) the traffic information associated for each year. The economic importance of these highways which is a part of the National Highway System makes these routes more critical with more possibility of crashes to occur every year (Ye et al., 2013).

To facilitate the crash data collection, the freeway has been divided into 274 roadway segments. These segments vary in lengths with roughly 0.87 miles mean segment length and 0.60 mile standard deviation. These segments were interchange segments, with interchange segments defined as segments bounded by the farthest ramp terminal on either side of an interchange overpass. A noteworthy point to be consider is that all the analyzed segments in this thesis are interchange segments. Interstate (5) has some complex interchanges but on average, the definition of interchange only segment means

the length between farthest ramp points on either side of overpass. Thus, the lengths of these segments would be far less. 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. The crashes sample size produces 822 (=274x3) segment-year observations.

Figure (3-1) Interstate highway (no. 5) in Washington state USA.

In total, 13,359 individual crashes were included in this study. Crash frequencies by type and year are shown in Table (3-1).

Table (3-1) Crash frequencies by type and year

Year	Total accidents	Crash type						
		Rear-End	Sideswipe	Fixed objects	Overturn	Other types: Same Direction	Head-on	Others
2005	4,550	2,578	775	684	79	228	6	199
2006	4,519	2,543	785	683	82	265	1	160
2007	4,290	2,391	817	637	81	192	5	166
∑	13,359	7,512	2,377	2,004	242	685	12	525
		56.2%	17.8%	15.0%	1.8%	5.1%	0.1%	3.9%

On state highways, the major collision types usually are: same direction

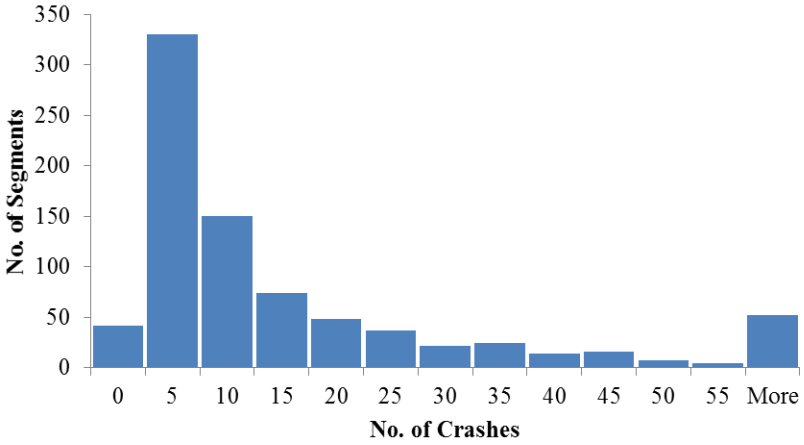
collisions resulting in rear end or sideswipe crashes, as well as fixed objects and entering at angle. The proportion of entering at angle collisions on interstates, especially on the mainline is virtually negligible, with the result that three major collision types dominate the frequency distribution. This is confirmed by Table (3-1), which shows that rear-end, sideswipe, and fixed objects crash types comprise a large proportion of the frequency distribution of crash types, with the remaining category classified as “all other,” and inclusive of same direction collisions not resulting in rear end or sideswipe crashes; head-on; crashes involving vehicle fire, pedestrians, parked vehicles, wrong way crashes, etc.

3.2 CRASH DATA DISTRIBUTIONS AND DESCRIPTIVE STATISTICS

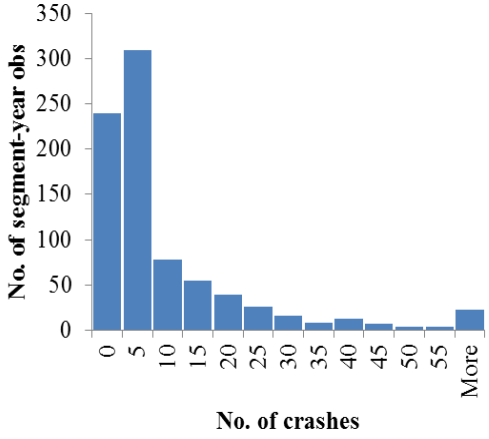
The distribution of crash frequencies among segments per year by each crash type is shown in Figure (3-2). (Zero counts are 30%, 12%; 11%; and 19% for rear end, sideswipe, fixed objects and all-other respectively.) Zero-inflated count models were not considered in this study. For traffic and geometric characteristics, a data from Washington State Department of Transportation (WSDOT, http://www.wsdot.wa.gov/) databases were used related to each segment on the roadway. Geometric data include percentage of lanes cross section proportion by length of segment, maximum and minimum radii of horizontal curves, central angle of horizontal curves, maximum grade, minimum grade, grade differential, tangent length, and number of changes in grade for vertical curves, number of horizontal curves per segment, number of vertical curves per segment, presence of interchanges and presence of exit and entrance ramps. Traffic operations data include average annual daily traffic. It should be noted that since the sample size is limited to three observations from the same segment (number of crashes per year) are treated as independent in this study although significant serial correlations are anticipated. The consideration of this correlation remains as a future research task.

Table (3-2) provides information on the mean and standard deviation of selected variables in the dataset. There are basically twelve exogenous variables representing the explanatory variables for traffic volume and

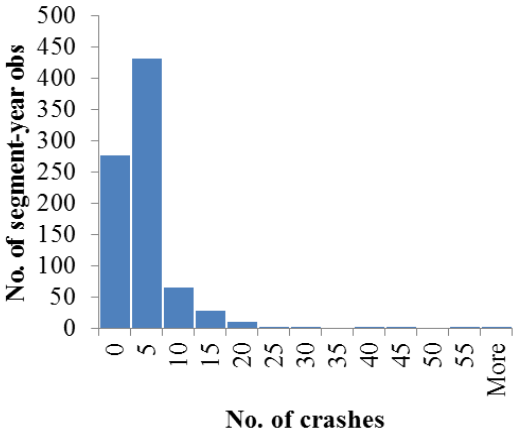
geometrics of each segment on the freeway, and logarithmic conversions of annual average daily traffic volume and length in miles are used in the model. Four endogenous variables represent the crash types are considered in this study.



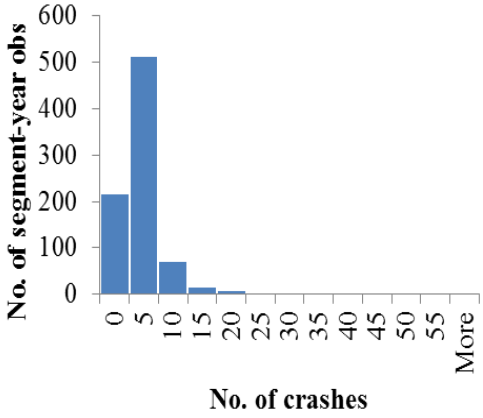
(a) Total crash count



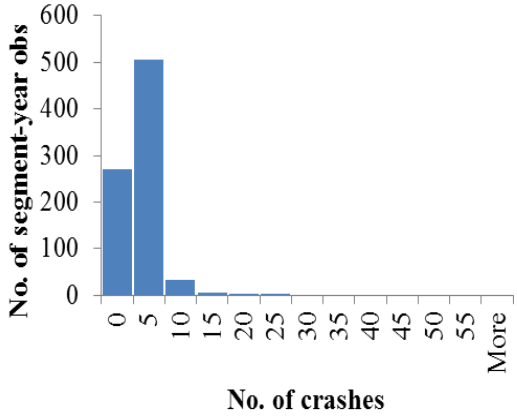
(b) Rear end



(c) Sideswipe



(d) Fixed object



(e) All-other

Figure (3-2) Distribution of crash frequency by type

Table (3-2) Descriptive statistic of observed crash count and explanatory variables

	Explanatory Variables	Mean	St. dev.	Max	Min
Crash types	Number of rear-end accidents per year	9.14	20.44	212.00	0.00
	Number of sideswipe accidents per year	2.89	5.71	65.00	0.00
	Number of fixed object accidents per year	2.44	3.32	33.00	0.00
	Other Types [Same direction, overturn, head-on, others]	1.78	2.42	21.00	0.00
	Total number of accident count per year (sum of all types of accidents record)	16.25	29.70	305.00	0.00
Explanatory variables	Annual average daily traffic volume in vehicles per hour (AADT)	17207.80	7151.08	42214.18	1079.49
	Logarithm of AADT	9.65	0.51	10.65	6.98
	Length in miles	0.87	0.60	4.13	0.13
	Logarithm of segment length	-0.30	0.55	1.42	-2.04
	Urban rural dummy, 1 if rural, 0 if urban	0.27	0.44	1.00	0.00
	Percentage of three lanes or larger {up to 5 lanes} cross section proportion by length of segment.	0.66	0.47	1.00	0.00
	Number of horizontal alignments per segment.	1.57	1.51	13.00	0.00
	Interchange type dummy for diamond ramps	0.49	0.50	1.00	0.00
	Largest vertical curve rate of vertical curvature in segment	1.20	1.23	5.79	0.00
	Shortest vertical curve length in segment in miles	0.08	0.07	0.45	0.00
	Smallest vertical curve rate of vertical curvature in segment	847.75	1,836.61	20,000.00	0.00
	Largest beginning vertical curve elevation in segment	-1.75	142.47	194.72	-333.71
	Largest horizontal curve central angle in segment	15.22	15.21	9.24	0.00
	Number of vertical curves in segment	2.59	2.00	13.00	0.00

Finally, the presence of crash type dependency is shown in Figure (3-3), were each pair exhibits a heavy scattered left tail.

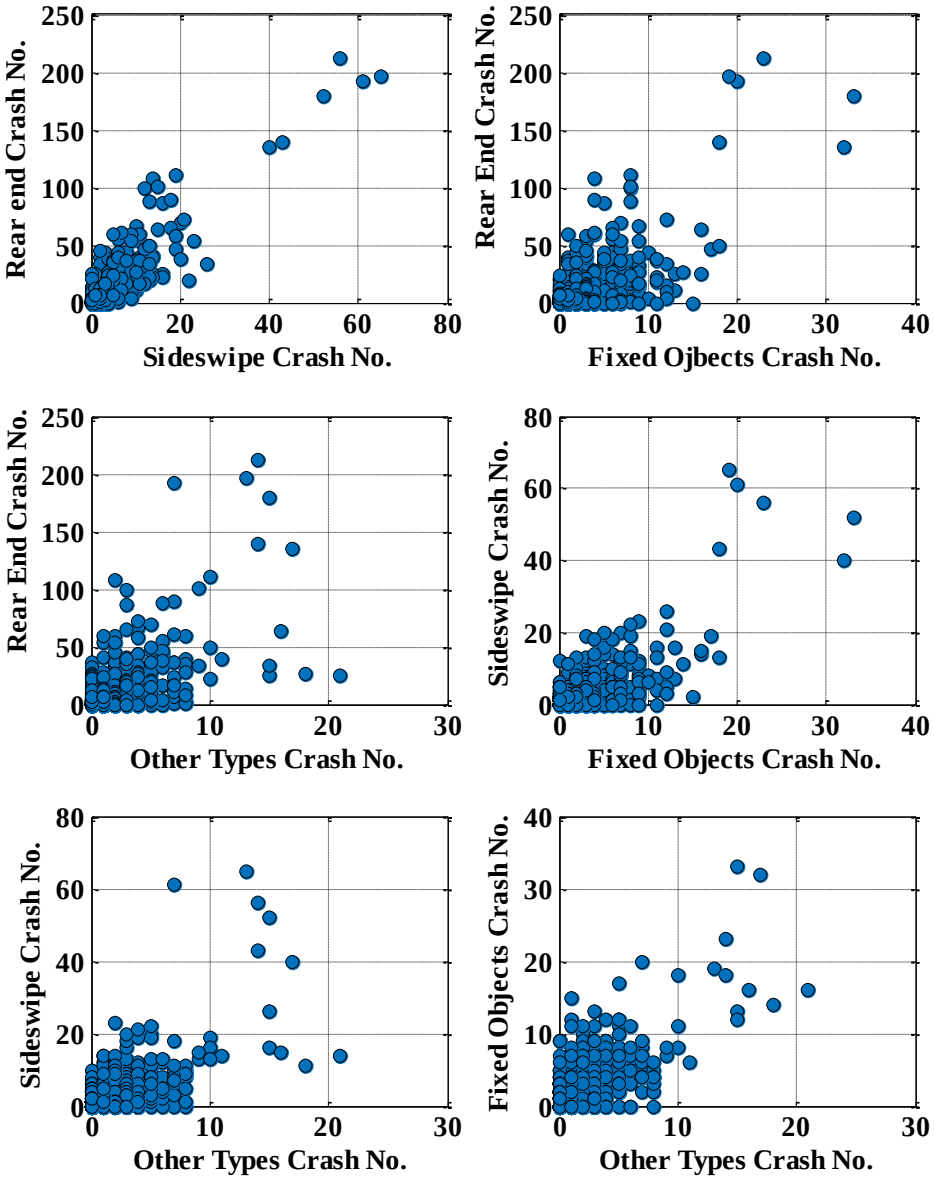


Figure (3-3) Crash type scatter plots

Chapter 4

Multivariate Poisson Gamma Mixture Model

4.1 INTRODUCTION

This Chapter investigates the correlations and covariances among the rear end, sideswipe, fixed object and other crash types on freeway sections using three-year crash data for 274 multilane freeway segments in the State of Washington, U.S.A. A multivariate Poisson gamma mixture count model (MVPGM) is developed assuming positive correlation among crash types. The model parameters are estimated using a maximum likelihood approach. The objective of the proposed model is to investigate if the unobserved heterogeneity correlations among different types of crash frequencies are significant or not. In addition to evaluating crash type correlations and covariances by crash type, the model also allows for evaluation of roadway geometric marginal effects and how they compare with crash type-specific effects.

4.2 BACKGROUND

An enormous body of literature has been devoted to modeling crash and safety considerations. The concept of multivariate frequency modeling that considering the error term correlations has been incorporated for multiple aims in the field of transportation. This concept is exploited to address the crash modeling with error components that represent the unobserved heterogeneity by jointing the probability of more than one crash type could occur on a specific segment of the freeway.

Basically, there are five multivariate count models to estimate the correlation among frequencies by crash types or severity: multivariate Poisson model; multivariate negative binomial model; multivariate Poisson-gamma

mixture model; multivariate Poisson-log-normal model and latent Poisson-normal model. Multivariate Poisson-log-normal models have been used extensively in the literature for both crash types and severity (see for more details, Winkelmann, 2008). Ye *et al*. (2013) modeled crash frequency by severity at freeway using simultaneous equations Poisson log normal model with an error component structure that are normally distributed. Chiou and Fu (2013) also modeled the crash frequency by severity using multinomial-generalized Poisson model with error components. In these two studies, the maximum simulated likelihood approach has been adopted in order to solve the integral of the conditional joint probability. Both of these studies also have used the normal distribution to model the error component structure for each individual event. Park and Lord (2007), Basyouny and Sayed (2009), Ma et al. (2008) and Ma et al. (2006) have used the Bayesian approach. Nevertheless, “there is some doubt whether these models could be time consuming when applied to high dimensional multivariate data” (Winkleman, 2008). Beside this fact, Bayesian approach doesn't tell a correct way to select a distribution of the prior; the posterior is heavily influenced by this selection and finally the simulations provide slightly different answers unless the same random seed is used (Winkelman, 2008). The other model is multivariate latent Poisson-normal that proposed the non-linear parameterization of the thresholds as a function of exogenous variables (Castro et al., 2013; Castro et al., 2012). Complexity and non-closed-form for the joint probability are considered as the main hindrance to estimate parameters of these models.

The Poisson gamma mixture model was first introduced by Hausman *et al.* (1984) with further explanation of its use by Dey and Chung (1992). In the Hausman-et al/Miles models, correlation is generated by an individual specific multiplicative error term. Miles (2001) provides an application of this model to individual consumer data on the number of purchases of bread and cookies in one-week period, using maximum likelihood estimation (MLE) of the Poisson gamma mixture probability. Kockelman (2001) conducted a time and budget constrained activity demand analysis utilizing the same model as Miles (2001). This model offers a closed form and is easy to estimate by using the maximum likelihood method. Thus, this chapter develops a multivariate Poisson gamma

mixture (MVPGM) model to simultaneously model crash frequencies by type considering the effects of various roadway, geometric and traffic volume factors on crash frequencies. Moreover, the proposed model considers the covariance matrix through error components specified under an integrated model framework among crashes types. Model estimation is achieved through the use of maximum likelihood estimation (MLE) method that provide consistent and efficient parameter estimates, and test statistics for hypotheses testing.

The rest of the chapter is structured as follows. The next section provides the empirical crash types and the associated explanatory variables. Section three offers the methodology of the proposed count model while section four is the application of this model to our crash count data. Finally, the fifth section is the conclusions that we draw from this application with a direction of the future

4.3 EMPIRICAL CRASH DATA SETTING

As we mentioned in Chapter 3, the crash dataset is obtained for interstate 5 in the State of Washington, USA. Three years of crash data were collected from 2005 to 2007. Data contained three different categories: The crash-record here is considered as a cross-sectional count data, which means the observation period is not considered here. Thus, the crash-count sample size produces 822 ($=273\times3$) segment-year observations. Four crash types we will considered in developing our model in this chapter. Rear-end, Sideswipe, fixed object and 'all-other' types. The crash-count type distributions are presented in Figure (3-2) while the descriptive statistic of the main explanatory variables in this study is shown in Table (3-2). Selection of the explanatory variables for each crash type was a results of conducting several univariate NBII model regression. We will allow for each crash type to have a different set of configurations for the associated parameters, even-though we will maintain almost same explanatory variables among these crash types.

4.4 MODEL SPECIFICATIONS

4.4.1 Selection of the count model

As already stated, this chapter deals with constructing a multivariate Poisson-gamma mixture (MVPGM) model to analyze the crash types where rear-end, sideswipe, fixed objects and other types are considered. Meanwhile, the correlations among these types are taken into account.

At first, the univariate negative binomial (NB) regression model is utilized. This model is widely used in counts data applications due to its simplicity and commonly used for modeling crash frequencies (Cameron and Trivedi, 1986). Notwithstanding the purpose of this paper is to develop a multivariate type of crash frequencies model, the univariate NB model is presented for three purposes, first is to investigate the overdispersion problem when the expected variance is larger than the expected mean which is the most common issue for crash data (Mannering and Bhat, 2014). Second is to assist our selection of the most significant explanatory variables related to each crash type. Finally, it would be used as a reference to compare with the proposed model.

4.4.2 Univariate NB model

Univariate NB model can be derived as Poisson-gamma mixture and given as

$$\begin{aligned}\Pr(y_{ji}) &= \int \left[\frac{\exp(-\lambda_{ji}u_j)\times(-\lambda_{ji}u_j)^{y_{ji}}}{\Gamma(y_{ji}+1)} \right] g(u_{ji})du_{ji} \\ &= \frac{\Gamma(y_{ji}+\theta_j)}{y_{ji}!\Gamma(y_{ji})}\left(\frac{\lambda_{ji}}{\lambda_{ji}+\theta_j}\right)^{y_{ji}}\left(\frac{\theta_j}{\lambda_{ji}+\theta_j}\right)^{\theta_j}\end{aligned} \tag{4-1}$$

where y_{ji} is the number of type j crash on the segment i, $g(u_{ji})$ is the gamma mixture function of the dispersion parameter, θ_j is dispersion parameter, and λ_{ji} is the expectation of number of type j crash on segment i and given as

$$\lambda_{ji} = \exp(\mathbf{x}'\beta_j + u_{ji}) \tag{4-2}$$

where x_i is a vector of explanatory variables, and β_j is a weight vector of each explanatory variable. The univariate NB model for total crash number is similarly given (by substituting j to T in Eq.(4-1) and Eq. (4-2) where T is used as an index for total crash).

4.4.3 MVPGM model specification

The modeling of the correlation structure in multivariate count data is vigorously urgent in enhancement the parameter estimates efficiency which patronizes the correction of the standard errors computations (Winkelmann, 2008). The Poisson-gamma mixture model (Cameron and Trivedi, 1986; Mannering and Bhat 2014) can be generalized and extended to allow for unobserved heterogeneity and overdispersion in the respective marginal distributions. The proposed model for here (see Hausman *et al.* 1984 for details) incorporates a mixture multivariate density for expected crash y_{ji} (the dimensions can vary depending on number of different crash types) which is obtained after integrating out the segment specific heterogeneity, which is assumed to be common across crash types due to unobserved effects that are cross sectional for segment *i*. The mixture multivariate density of expected crash (y_{ji} , …, y_{Ji}) for *J* dimensions is obtained after taking integral.

$$\Pr(y_{1i},\ldots,y_{Ji}) = \int \left[\prod_{j=1}^{J} \frac{\exp(-\lambda_{ji}u_i)(-\lambda_{ji}u_i)^{y_{ji}}}{\Gamma(y_{ji}+1)} \right] g(u_i)\,du_i \tag{4-3}$$

where $g(u_i)$ is the mixture function of the error term u_i . Let u_i follows a gamma distribution with E(u_i) =1 and Var(u_i) = θ_M^{-1} property, at this instant, constructing the joint distribution function of y_i (among crash types) is given as a solution of the mixture function above, which leads to obtain the negative binomial distribution function (Winkleman, 2008) as

$$\begin{aligned} \Pr(y_{1i},\ldots,y_{Ji}) &= \left(\prod_{j=1}^{J} \frac{(\lambda_{ji})^{y_{ji}}}{\Gamma(y_{ji}+1)} \right) \frac{\theta_M^{\theta_M}}{\Gamma(\theta_M)} \int e^{-u_i(\lambda_{Ti}+\theta_M)} u^{y_{Ti}+\theta_M-1}\,du_i \\ &= \left(\prod_{j=1}^{J} \frac{(\lambda_{ji})^{y_{ji}}}{\Gamma(y_{ji}+1)} \right) \frac{\Gamma(y_{Ti}+\theta_M)}{\Gamma(\theta_M)} \theta_M^{\theta_M} (\lambda_{Ti}+\theta_M)^{-(y_{Ti}+\theta_M)} \end{aligned} \tag{4-4}$$

where y_{Ti} is the sum of the number of crashes of different types, and equal to the number of total crashes.

It is worth mentioning that the univariate Poisson-gamma count model is mathematically similar to the proposed model. The difference is only in the aspects of how to deal with the unobserved heterogeneity that lead to drive these two models. As for the Poisson-gamma model the u_i is used rather than

u_{ji} as suggested in the MVPG model (Winkleman, 2008). The covariance among the crash types for a given segment *i* is given by

$$\begin{aligned} Cov(y_{ji}, y_{ki}) &= E_u Cov(y_{ji}, y_{ki} \mid u_i) + Cov_u\left[E(y_{ji} \mid u_i), E(y_{ki} \mid u_i) \mid u_i\right] \\ &= 0 + Cov_u\left(\lambda_{ji} u_i, \lambda_{ki} u_i\right) \\ &= \theta_M{}^{-1} \times \lambda_{ji} \times \lambda_{ki} \end{aligned} \tag{4-5}$$

where j, k: index of the type of crash, E(.): expected value; and Cov(.): Covariance for the selected pair of crash types

One of the distinct features of this model is that the covariance is not equal across the count outcomes. Rather, it is totally reliant on the product of the expected values of $\lambda_{ji}\lambda_{ki}$. This feature is useful for modeling the nonnegative random variables. The restriction in the upper bound will be eliminated in that case which is a common deficiency associated with the other types of multivariate count models. We can write the unconditional correlation between two count variables as follows

$$Cor(y_{ji}, y_{ki}) = \frac{\lambda_{ji}\lambda_{ki}}{\sqrt{\left(\theta_M \lambda_{ji} + \lambda_{ji}{}^2\right)\left(\theta_M \lambda_{ki} + \lambda_{ki}{}^2\right)}} \tag{4-6}$$

One might think of a possible disadvantage of the MVPGM model is that the covariance and the dispersion are not estimated separately. Thus, it might mislead to judge whether a significant theta could be either a sign of over-dispersion occurrence or due to the correlation among the count outcomes (or both) (Winkleman, 2008). Rewriting the joint probability in Eq.(4-4):

$$\Pr(y_{1i}, \ldots, y_{Ji}) = \frac{\Gamma(y_{Ti} + \theta_M)}{\Gamma(\theta_M)\prod_{j=1}^{J} y_{ji}!}\left(\frac{\theta_M}{\lambda_{Ti} + \theta_M}\right)^{\theta_M} \prod_{j=1}^{J}\left(\frac{\lambda_{ji}}{\lambda_{Ti} + \theta_M}\right)^{y_{ji}} \tag{4-7}$$

The log-likelihood function across all segments *i* would be,

$$\Pr(y_{1i},...,y_{Ji}) = \sum_{i=1}^{N}\left\{\left[\log\Gamma\left(\sum_{j=1}^{J}(y_{ji})+\theta_M\right)-\sum_{j=1}^{J}\log(y_{ji}!)-\log(\Gamma(\theta_M))\right]\right.$$
$$+\sum_{j=1}^{J}\left[y_{ji}\times\left(\log(\lambda_{ji})-\left(\log\left(\sum_{j=1}^{J}\lambda_{ji}\right)+\theta_M\right)\right)\right] \quad (4\text{-}8)$$
$$\left.+(\theta_M\times\log(\theta_M))-\left(\theta_M\times\left(\log\left(\sum_{j=1}^{J}\lambda_{ji}\right)+\theta_M\right)\right)\right\}$$

The Log Likelihood function was coded using GAUSS programming language (Aptech 1999). We utilized the BFGS algorithm which is offered by the maxlik library to maximize this function. As we mentioned before, the count crash data were assumed to be a cross-sectional data where the time is irrelevant

4.5 MODEL ESTIMATION AND PERFORMANCE

4.5.1 Estimation Results of Univariate NB Model

To begin with, a univariate NB model of the total number of crashes was estimated as a function of exogenous variables. With the same specification, four separated univariate negative binomial models for each crash type were estimated, the results of which are shown in Table (4-1). The value of θ is estimated to be 3.672 in the NB model for total number of crashes. This parameter is statistically significant as evidenced by the larger t-value. The value of θ for each crash type are 1.690 for rear end, 4.560 for sideswipe, 3.528 for fixed object and 4.759 for the all-other types. The univariate models were also tested for the plausibility of zero-inflated specifications (Shankar et al., 1997). The Vuong statistics for these models were strong negative values (<-4.00) suggesting that the preferred models were the baseline negative binomial specification. The Vuong statistics were tested for zero-inflated specifications where the zero-state had the same vector of covariates as the count state. The univariate NB model variables were used as the specification set for the estimation of the multivariate Poisson-gamma model discussed below.

4.5.2 Estimation Results of MVPGM Model

Estimation results of the MVPGM regression model with correlation are presented in Table (4-2). The estimation results provide parameter estimates for four types of crashes. The estimated dispersion parameter is estimated to

be 3.654, which implies an overdispersion magnitude of 0.274. This parameter is statistically significant as evidenced by the larger t-value. The log-likelihood is -6,310.50 while AIC and BIC are 12,681.10 and 12,822.40 respectively. The combined log-likelihood value of the univariate crash type models (from Table (4-1)) is found to be -6,079.20. The result indicates the separate univariate crash type models fit better than the proposed model. The possible reason is that the differing magnitudes of overdispersion in the univariate models are ignored in the multivariate model. In the multivariate model, the overdispersion parameter is assumed dot be the common for the four crash types. Compared to the value of 3.654 for the MVPGM model, the univariate model values of θ are 1.690, 4.560, 3.528 and 4.759 for rear end, sideswipe, fixed object and all-other types respectively. The value is statistically different form one another, implying the different magnitudes of unobserved heterogeneity for different crash types.

The three variables commonly significant in all crash type functions include: ADT, segment length and lane cross section proportion (3 lanes or grater). The number of horizontal curves variable and the diamond interchange dummy are significant in the rear end, sideswipe and fixed object models. The effect of the horizontal curves parameters is positive indicating that as the number of curves increases, the expected number of rear end, sideswipe and fixed object crashes will increase as well. The contributing factors arising from horizontal curvature appear to be multifaceted – from speed differentials and their associations with rear ends, to, lane offsetting and sideswipes, and potential loss of control and roadside encroachments resulting in fixed object collisions. The diamond interchange dummy has a negative effect due to the fact that diamond interchange Footprints are typically larger and provide for adequate merge and weave distances, thereby decreasing the likelihood for rear end, sideswipe or fixed object crashes. The urban-rural dummy is significant in rear-end and sideswipe crah functions indicating that in rural contexts, the expected number of rear-end and sidewise collisions is not as high as it would be in urban contexts- an outcome of congesting related effects. The urban/rural classification is not based on crash reports. Crash report classification of crash location as rural or urban is inconsistent. So, if we were

to use crash reports as a basis, we would get an inconsistent aggregation of counts. For this reason, we used the urban/rural classification based on the State Highway Log provided by the Washington State Department of Transportation. This definition is based on population levels (urban and urbanized areas with populations exceeding at least 50,000; or rural). Vertical curvature variables are sporadically significant – with the minimum vertical grade variable being significant in the sideswipe model, while the maximum vertical curve elevation variable is also significant in the same model. The horizontal curve central angle variable appears to be significant in the fixed object model.

It is found that the smallest absolute vertical curve gradient in segment is associated positively with sideswipe crash type frequency. Three likely scenarios can occur where the smallest absolute value vertical curve gradient can become influential. In the first scenario, if all curve gradients in a segment are positive, the smallest positive gradient represents the smallest upgrade in the segment. In the second scenario, if all curve gradients in a segment are negative, then, the smallest curve gradient represents the smallest downgrade in the segment. In the final scenario, if there is a mixture of positive and negative gradients in the segment, then, the smallest absolute gradient represents either the smallest upgrade or smallest downgrade in the segment. In all three cases, the effect reflects the impact of the smallest grade whether it is an upgrade or downgrade. Segments where vertical curves are not present were categorized as flat segments and therefore implicitly serve as the baseline or the smallest absolute vertical gradient parameter. Our finding notes that as the smallest grade increases, its effect likely results in greater speed differentials, compared to segments with no vertical grades. One would have expected the steepest absolute gradient to be influential, but it turns out the smallest grade is statistically significant. This in turn is likely to result in increased lane changing frequencies which can increase the likelihood of sideswipe crashes. Another variable associated with the sideswipe crash type is the largest beginning vertical curve elevation in segment. This variable is found to be statistically significant at the 1 percent level and of significant magnitude (0.743). When a leading vehicle reduces its own speed due to a

high elevation of the vertical curve, the following vehicle tries to pass the leading vehicle raising the likelihood of lane changing due to speed differentials. Similar to the above-mentioned effect of curve gradient, the likelihood of sideswipe crashes can increase as the maximum curve elevation in segment increases.

Table (4-3**)** shows the correlations among crash types for a given segment as calculated using Eq.(4-6). A restriction of MVPGM is given by the fact that it constrains the correlation among counts to be positive (Gurmu and Elder, 2000). The correlations range between 0.399 and 0.299, which demonstrate the presence of common unobserved factors that affect crash type frequency. These common unobserved factors that influence crash frequencies by crash type include pavement condition, environmental and weather conditions, driver population factors, adjacent land use characteristics, traffic composition variables (trucks, buses, etc.), sight distance and others (Ye *et.al*., 2009).

Table (4-1) Univariate NB models of total and type specific crashes

Explanatory Variables	Total crashes		Rear End		Sideswipe		Fixed Object		All-other	
	Coefficient	t-stat.	Coefficient	t-stat.	Coefficient	t-stat.	Coefficient	t-stat.	Coefficient	t-stat.
Constant	-12.957**	-11.50	-23.380**	-10.60	-16.327**	-13.67	-6.998**	-4.87	-11.140**	-11.77
LnAADT	1.545**	13.37	2.488**	10.95	1.705**	13.92	0.800**	5.62	1.180**	11.98
LnLength	0.759**	10.40	0.559**	5.45	0.708**	7.79	0.885**	9.15	0.997**	16.97
Urban rural dummy, 1 if rural, 0 if urban	-0.326**	-4.25	-0.665**	-4.54	-0.717**	-5.51	-0.199*	-1.72	-	
Proportion of three or more lanes cross section by length of segment	0.750**	12.58	0.961**	9.58	0.611**	6.16	0.360**	4.02	0.457**	5.05
Number of horizontal curves per segment	0. 102**	4.14	0.154**	4.34	0.118**	4.57	0.070*	2.51	-	
Diamond interchange type dummy	-0.286**	-5.01	-0.269**	-3.02	-0.281**	-3.71	-0.241**	-2.83	-	
Smallest vertical gradient in segment	0.113**	3.88	-		0.072**	2.74	-		-	
Shortest vertical curve length in segment in miles	-2.123**	-3.17	-		-		-		-	
Largest vertical curvature rate in segment	-0.028**	-3.23	-		-		-		-	
Largest beginning vertical curve elevation in segment	-		-		0.884**	2.66	-		-	
Largest horizontal curve central angle in segment*	-		-		-		0.496*	2.55	-	
Number of vertical curves in segment	-		-		-		-0.049*	-1.98	-	
θ_j	3.672**	14.34	1.690**	13.74	4.560**	7.83	3.528**	6.77	4.759**	5.31
Sample size	822		822		822		822		822	
Log-likelihood at convergence	-2,505.50		-1,976.10		-1,350.80		-1,474.20		-1,278.10	
AIC	5,033.00		3,968.10		2,721.60		2,968.30		2,566.20	
BIC	5,084.90		4,005.80		2,768.70		3,015.40		2,589.80	

- Not relevant; ** Significant at 1% level; * Significant at 5% level. Significance of the actual overdispersion parameter (1/θ) is estimated using the delta method. Significance of the overdispersion effect is very strong for all crash types, at or better than the 99.5% level.

Table (4-2) MVPGM model of four types of crashes

Explanatory Variables	Rear End		Sideswipe		Fixed Object		Other Types	
	Coefficient	t-stat.	Coefficient	t-stat.	Coefficient	t-stat.	Coefficient	t-stat.
Constant	-19.356**	-11.84	-16.666**	-12.12	-7.162**	-4.27	-10.930**	-9.74
LnAADT	2.074**	12.43	1.743**	12.34	0.809**	4.78	1.160**	9.94
LnLength	0.617**	6.23	0.673**	7.23	0.781**	9.09	0.904**	15.69
Urban rural dummy, 1 if rural, 0 if urban	-0.844**	-5.21	-0.659**	-5.05	-0.177	-1.48	-	
Proportion of three or more lanes cross section by length of segment	1.137**	12.05	0.554**	5.64	0.338**	3.63	0.411**	4.48
Number of horizontal curves per segment	0.153**	4.77	0.117**	4.20	0.068*	2.54	-	
Diamond interchange type dummy	-0.347**	-4.31	-0.248**	-3.22	-0.174*	-2.13	-	
Smallest vertical gradient in segment	-		0.051*	2.37	-		-	
Shortest vertical curve length in segment in miles	-		-		-		-	
Largest beginning vertical curve elevation in segment	-		0.743*	2.43	-		-	
Largest horizontal curve central angle in segment*	-		-		0.425*	2.17	-	
Number of vertical curves in segment	-		-		-0.032	-1.38	-	
θM	3.654** (14.80†)							
Sample size	822							
Log-likelihood at convergence	-6,310.50							
AIC	12,681.10							
BIC	12822.4							

- Not relevant; ** Significant at 1% level; * Significant at 5% level.
†t-stat. of the θM.

Table (4-3) Estimated correlation matrix for a given segment

	Rear-End	Sideswipe	Fixed Object	Other Types
Rear-End	1.000			
Sideswipe	0.388 (20.130†)	1.000		
Fixed Object	0.399 (12.086)	0.328 (8.799)	1.000	
Other Types	0.362 (3.595)	0.299 (2.617)	0.316 (1.860)	1.000

† Covariance of the error term value are between parentheses

Eq.(4-5) is further utilized for the comparison of covariances resulting from the MVPGM and univariate models. To make this comparison, the observed covariance of crash type frequencies is used as the benchmark. The MVPGM covariance is the sum of two components; the covariance between the expected numbers of crash types and the covariance resulting from the unobserved heterogeneity given by Eq.(4-5). The covariance between expected numbers of cash types is a measure of observed heterogeneity captured via explanatory variables in the cash type models. The covariance calculated by the univariate models does not include the unobserved heterogeneity effects from Eq.(4-5) because the unobserved heterogeneities are assumed to be independent. The results of the analysis of covariance are presented in Table (4-4). The table shows that the MVPGM covariances of crash type are closer to the observed covariances for the corresponding crash type pairs. The results suggest that the MVPGM is a plausible way to capture the composition of the total covariance among crash type, in spite of the fact that the dispersion parameter is restricted to be the same across crash types.

Table (4-4) Covariance of the numbers of crashes between crashes types

Pair of crash types	Observed	MVPGM	Univariate models
Rear end and sideswipe	104.097	70.182	56.138
Rear end and fixed object	46.118	35.897	27.001
Rear end and other types	29.177	26.029	19.762
Sideswipe and fixed objects	13.730	10.251	7.685
Sideswipe and other types	8.885	7.430	5.562
Fixed objects and other types	5.304	4.542	3.282

Since the sum of the crash types is equal to the number of total crashes, it is desirable that the variance structures of the type specific models be consistent with that of the total crash model. An evaluation of the variance

structure in terms of the variance of the expected number of crashes versus variance of the number of crashes for a given segment can also provide useful insights. The variances calculated using the total crash model can be compared against the variances form the MVPGM and univariate models using variance decomposition as described below. The variance of the expected number of crashes for a segment can be calculated from the estimated type specific crash models (using the variance of sum of random variables formula) by

$$Var\left(\sum_{j=1}^{J}\lambda_j\right)=\sum_{j=1}^{J}Var\left(\lambda_j\right)+2\sum_{\{(j,k)|j<k\}}Cov\left(\lambda_j,\lambda_k\right) \tag{4-9}$$

The above formula will yield different values for MVPGM and univariate NB models depending on the estimate of the mean, which is likely to differ for univariate versus MVPGM models with the same specifications.

The variance of the number of crashes for a given segment can be evaluated from the estimated MVPGM through the relationship

$$\lambda_i+\theta_T^{-1}\lambda_i^2=\sum_{j=1}^{J}\left(\lambda_{ji}+\theta_M^{-1}\lambda_{ji}^2\right)+2\sum_{\{(j,k)|j<k\}}\left(\theta_M^{-1}\lambda_{ji}\lambda_{ki}\right) \tag{4-10}$$

where θ_T: dispersion parameter of total crash model, and θ_M: dispersion parameter of the MVPGM model

Since the covariances of crash types do not factor into the univariate models, the variance of the number of crashes is given by

$$\lambda_i+\theta_T^{-1}\lambda_i^2=\sum_{j=1}^{J}\left(\lambda_{ji}+\theta_j^{-1}\lambda_{ji}^2\right) \tag{4-11}$$

where θ_j: dispersion parameter of the designated univariate NB model for crash type j.

Table (4-5) presents the results of the variance analysis based on the application of Eq.(4-9), Eq.(4-10), Eq.(4-11). Similar to the insights form Table (4-4), the table shows that both the variances of the MVPGM model are in closer agreement with those of the total crash model. This finding applies to the variance of the expected number, as well as variance of the count. Table (4-4) and Table (4-5). Combined indicate that the MVPGM model capture variance

and covariance structures in close agreement with those produced by total crash models.

Table (4-5) Variance structure of total crashes

	Total crash model	MVPGM	Univariate Models
Variance of the expected number of crashes among segments	403.22	414.39	474.81
Variance of the number of crashes for a given segment	189.40	194.02	197.39
Sum	592.62	608.42	672.20

4.5.3 Marginal effects

The marginal effect of the l^{th} explanatory variable x on dependent variable y can be expressed by taking the first derivative of the expected number of type specific crashes estimated by the MVPGM model, it can be expressed as,

$$\frac{\partial E\left(y_{ji} \mid x_{ji}\right)}{\partial x_{jl}} = \beta_{jl} \exp\left(x'_{ji}\beta_j\right) \tag{4-12}$$

The marginal effect calculated by the total crash model is shown in column 1 in Table (4-6) while the marginal effects for type specific crashes calculated by the MVPGM model are shown in columns 2-5. It is noticeable that the marginal effects of rear-end crashes are larger in the absolute value than any other crash types consistently for all explanatory variables. The results suggest that interstate rear-end crash likelihood is most sensitive to geometric and traffic conditions. In particular, the horizontal curve variable finding suggests that realignment of interstate segments might provide a larder rear-end reduction benefit than a fixed-object crash reduction benefit.

The last column of Table (4-6) represents the sum of the marginal effects of the four types of crashes calculated by the MVPGM model. The sums of the crash type marginal effects for Ln(AADT), Ln(Length), three-plus lane proportion, number of horizontal curve and diamond interchange dummy variables are similar in magnitude to the corresponding value for the total crash marginal effect.

Table (4-6) Marginal effects of total crash model and type specific crash models

Explanatory variable	Total No. of crashes	Rear-End	Sideswipe	Fixed Object	All-other	Sum
LnAADT	23.746	17.836	4.773	1.917	2.015	26.541
LnLength	11.670	5.285	1.842	1.850	1.571	10.548
Urban rural dummy, 1 if rural, 0 if urban	-0.385	-1.326	-0.934	-0.194	-	-2.453
Proportion of three or more lanes cross section by length of segment	11.532	9.782	1.516	0.801	0.714	12.814
Number of horizontal curves per segment	1.574	1.319	0.322	0.162	-	1.802
Diamond interchange type dummy	-4.397	-2.985	-0.678	-0.413	-	-4.076
Smallest vertical gradient in segment	1.738	-	0.141	-	-	0.141
Shortest vertical curve length in segment in miles	-32.632	-	-	-	-	0.000
Largest vertical curvature rate in segment	-0.429	-	-	-	-	0.000
Largest beginning vertical curve elevation in segment	-	-	2.034	-	-	2.034
Largest horizontal curve central angle in segment*	-	-	-	1.007	-	1.007
Number of vertical curves in segment	-	-	-	-0.075	-	-0.075

The results suggest the marginal effect of the total crash frequency could be accurately divided in to the marginal effects of crash type via the MVPGM model for the various explanatory variables as described above. This pattern is not observed for the urban-rural dummy, minimum absolute vertical gradient, shorter vertical curve length, largest vertical curvature, largest vertical curve beginning elevation and number of vertical curve variables. This is quite clear for the vertical curvature variables, which appear to be significant in the crash type models versus the total crash model varying degrees of consistency. It is plausible that specification consistency for vertical curvature variables may not be achievable at the total crash level while decomposing into crash types. Vertical curvature effects may be playing a nonlinear role in the development of speed differentials thereby creating apparent discordance in specification at the crash type level. Perhaps, functional form for vertical curvature needs to be researched at greater depth in order to achieve consistency in specification for total crash versus crash type models.

4.6 SUMMARY

This chapter has shown via a comparative analysis of univariate and multivariate models of crash types that the multivariate Poisson-gamma model appears to capture covariances across crash types consistent with the covariance of the total crash model. However, it also appears that the marginal effects of variables in their sum effect across crash types are not consistent with the marginal effect of the total crash model, especially for vertical curvature variables. This indicates that vertical curvature in its studied form (as a linear untransformed predictor) may be a poorly specified effect – a factor that deserves further research. Second, it appears that rear end crashes on interstates are most sensitive to geometrics and traffic volume, compared to other crash types, indicating the emphasis in the state of practice for active traffic management strategies. Third, it appears that further analysis in to the decomposition of variances by crash severity might lend an additional dimension to the understanding of the behavior of the dispersion parameter in joint models. While our model of crash types assumed a common dispersion parameter, the restriction that overdispersion be the same across crash types

manifested in poorer likelihoods in comparison to the univariate models which accommodated unrestricted dispersion parameters. This is somewhat of a significant finding – it is unclear whether the overdispersion restriction is causing likelihood convergence issues in the joint model, where premature convergence is indicated via a more negative log likelihood; or, if indeed, the likelihood is correct, then the impact of a single restriction (the overdispersion parameter) on causing a significant drop in the joint likelihood warrants further investigation via other datasets. A similar problem can also occur in the analysis of severities, where a similar likelihood discordance can occur between univariate versus multivariate analysis of severities. Nevertheless, there work here underscores the importance of seeking a methodological consensus for analyzing crash decomposition by type (crash type and severity included.)

Chapter 5

Multivariate Copula-Based Count Model

5.1 INTRODUCTION

Multivariate count model is developed by introducing a simple and practical formula in this chapter. 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. Our objective here is to develop a model which allows both positive and negative dependency among the count outcomes as well as offers a variety of dependent structures including radially asymmetric or tail dependency without a need for a simulation mechanism.

5.2 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. (2014b) 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 chapter, 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., 2014a) 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). 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; Shi and Valdex, 2014; and Nikoloulopoulos and Karlis, 2010).[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 Bhat et al., 2014c; Castro et al., 2013; Castro et al., 2012; Yamamoto and Morikawa; 2013; Sener et al., 2010; Ferdous et al., 2010; Paleti and Bhat, 2013). The CML approach is rooted in a general class of composite likelihood 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 Mannering et al., 2016; Mannering and

[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.

Bhat, 2014; Lord and Mannering, 2010 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) and Li et al. (2015). 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) 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. A characteristic of these studies is that 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 with CML estimation technique to model dependence across the observational unit. This approach provides flexibility in handling multiple marginals while taking advantage of the CML technique. Our contribution to the literature is to provide rigorous investigation methods on unobserved heterogeneous dependency across the types of crashes by copula-based approach including empirical copula diagnosis and investigations on the variance and covariance structure calculated from the estimation results of the multivariate count model.

The rest of this chapter is structured as follows. The next section provides the building blocks of the model in terms of formulation and inference. Section 5.5 describes the dataset including the selection of the crash types and the explanatory variables. Section 5.6 illustrates an application of the proposed model for analyzing the interstate crash types. The final section summarizes the important findings and conclusions from the study.

5.3 WHY COPULA?

The copula function is one of the available ways to construct the joint probability among different outcomes (continuous/discrete). From its basic definition, it can be considered as a multivariate cumulative function that leads usually to a tractable solution (see (Sklar (1959)). Precisely, the copula function binds together different outcomes and generate a multivariate distribution. The copula function merits lay in:

- Offer parametric distribution: in our multivariate crash-type counts context, negative binomial model is suitable to overcome the overdispersion problem.
- Offer non-parametric distribution: this point will be seen extensively in our study as a tool to determine which parametric copula should be used to model our crash-count data in Chapter 5 and Chapter 7 (empirical copula).
- Flexibility: most of available copula functions allow more flexible way to join different crash outcomes in a way to represent asymmetric-nonlinear correlations among the unobserved heterogeneity.
- Different way in parameterizing: for each marginal distribution selection there are several ways to parametrize the crash-count outcome. We will parameterize the expected crash-count (mean) parameter as a function of the explanatory variables, other possible way is to parameterize the correlation parameter of the copula itself.
- Time correlation representation: this feature can be seen in how the copula parameters based on the marginals can be pooled or allowed to vary across margins. This feature was the base for the model in Chapter 7.

These substantial features of the copula function will be exploited to cover each actuarial problem that we mentioned before. Table (5-1) shows the most popular copulas which are widely used in transportation engineering (see Bhat and Eluru, (2009) for more details).

Table (5-1) Most Popular parametric copula functions and their properties

Copula type	Function C(u1,u2)	Theta-Domain	Kendall's tau	Spearman's rho
Independent(Product)	$u_1 \times u_2$	Not applicable	0	0
Gaussian	$\Phi_G\left[\Phi^{-1}(u_1), \Phi^{-1}(u_2); \theta\right]$	$-1 \le \theta \le 1$	$\frac{2}{\pi}\arcsin(\theta)$ $-1 \le \tau \le 1$	$\frac{6}{\pi}\arcsin\left(\frac{\theta}{2}\right)$ $-1 \le \rho_s \le 1$
Clayton	$\left(u_1^{-\theta} + u_2^{-\theta} - 1\right)^{(-1/\theta)}$	$0 \le \theta \le \infty$	$\left(\frac{\theta}{\theta+2}\right)$ $0 \le \tau \le 1$	*No simple form* $0 \le \rho_s \le 1$
Frank	$-\frac{1}{\theta}\log\left(1+\frac{\left(e^{-\theta u_1}-1\right)\left(e^{-\theta u_2}-1\right)}{\left(e^{-\theta}-1\right)}\right)$	$-\infty \le \theta \le \infty$	$1-\frac{4}{\theta}\left[1-D_1(\theta)\right]$ $-1 \le \tau \le 1$	$1-\frac{12}{\theta}\left(D_1 \mid \theta\right)-\left(D_2 \mid \theta\right)-1$ $-1 \le \rho_s \le 1$
Gumbel	$\exp\left[-\left(\left(-\ln(u_1)\right)^{\theta}+\left(-\ln(u_2)\right)^{\theta}\right)^{(1/\theta)}\right]$	$1 \le \theta \le \infty$	$\left(1-\frac{1}{\theta}\right)$ $0 \le \tau \le 1$	*No simple form* $0 \le \rho_s \le 1$
Joe	$1-\left(\left(1-u_1\right)^{\theta}+\left(1-u_2\right)^{\theta}\right)^{(1/\theta)}$	$1 \le \theta \le \infty$	$1-\frac{4}{\theta}\left[1-D_1(\theta)\right]$ $-1 \le \tau \le 1$	*No simple form* $0 \le \rho_s \le 1$

$D_k(\theta_{ij})$ denotes to the Debye function given as $\frac{k}{\theta_{ij}^k}\int_0^{\theta_{ij}}\frac{t^k}{\left(e^t-1\right)}dt, k = 1,2.,\ u_1 = F_1(y_{1q})$ and $u_2 = F_2(y_{2q})$

5.4 METHODOLOGY

We begin with the formulation of a multivariate copula-based generalized ordered response count (MCORC) model of crash types. The basis is generalized ordered response (GOR) model in which symmetrical interdependence among the error terms of crash count types is assumed. This is extended later using the copula function to accommodate non-linear asymmetrical correlation.

5.4.1 Ordered Response Model with Count Data

Following the generalized ordered model representation in Castro et al. (2012), 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:

$$y_{iq}^{*}=\boldsymbol{\beta}_i'\mathbf{x}_{iq}+\varepsilon_{iq}, \qquad y_{iq}=k_{iq} \quad \text{if} \quad \eta_i^{k_{iq}-1}<y_{iq}^{*}<\eta_i^{k_{iq}} \tag{5-1}$$

where $\mathbf{x}_{iq}$ is a $(L\times 1)$ vector of non-intercept explanatory variables which are associated with a crash type, $\boldsymbol{\beta}'$ is a $(L\times 1)$ 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 $\beta'=0$, then we can write $y_{iq}^{*}=\varepsilon_{iq}$, which lead to,

$$\Pr\left(y_{iq}=k_{iq}\right)=\Pr\left(k_{iq}-1<y_{iq}\le k_{iq}\right)=\Pr\left(\eta_i^{k_{iq}-1}<\varepsilon_{iq}\le\eta_i^{k_{iq}}\right) \tag{5-2}$$

This relationship is essential to connect the continuous to the count distributions together. We can write Eq.(5-2) in terms of the cumulative density functions, as:

$$\begin{aligned}\Pr\left(y_{iq}\le k_{iq}\right)&=\Pr\left(\varepsilon_{iq}\le\eta_i^{k_{iq}}\right)\\&=F_i\left(k_{iq}\right)=H_i\left(\eta_i^{k_{iq}}\right)\end{aligned} \tag{5-3}$$

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

$$\eta_i^{k_{iq}}=\begin{cases}-\infty & k_{iq}=-1\\ H_i^{-1}\left[\Pr\left(y_{iq}\le k_{iq}\right)\right] & k_{iq}=0,1,....\end{cases} \tag{5-4}$$

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 a linear relationship as in the GOR model, instead, it follows the marginal distribution form congruously. Now we can write the multivariate cumulative probability density function $H_{(1,2,....,I)}$ for a given segment *q* as:

$$\begin{aligned}&\Pr\left(y_{1q}\le k_{1q},y_{2q}\le k_{2q},....,y_{Iq}\le k_{Iq}\right)=\Pr\left(\varepsilon_{1q}\le\eta_1^{k1q},\varepsilon_{2q}\le\eta_2^{k2q},....,\varepsilon_{Iq}\le\eta_I^{kIq}\right)\\&=H_{(1,2,....,I)}\left(\eta_1^{k1q},\eta_2^{k2q},....,\eta_I^{kIq}\right)\\&=\int_{-\infty}^{\eta_1^{k1q}}\int_{-\infty}^{\eta_2^{k2q}}....\int_{-\infty}^{\eta_I^{kIq}}f_{(1,2,....,I)}\left(\varepsilon_1,\varepsilon_2,....,\varepsilon_I\mid\boldsymbol{\Theta}\right)d\varepsilon_1\,d\varepsilon_2....d\varepsilon_I\end{aligned} \tag{5-5}$$

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

$$\Pr\left(y_{1q}=k_{1q},y_{2q}=k_{2q},....,y_{Iq}=k_{Iq}\right)= \int_{\varepsilon_1=\eta_1^{k_{1q}-1}}^{\eta_1^{k_{1q}}}\int_{\varepsilon_2=\eta_2^{k_{2q}-1}}^{\eta_2^{k_{2q}}}.....\int_{\varepsilon_I=\eta_I^{k_{Iq}-1}}^{\eta_I^{k_{Iq}}} f_{(1,2,....,I)}\left(\varepsilon_1,\varepsilon_2,....,\varepsilon_I \mid \boldsymbol{\Theta}\right) d\varepsilon_1\, d\varepsilon_2 .. d\varepsilon_I . \quad (5\text{-}6)$$

Eq. (5-6) has no closed form instead we have to evaluate the integral formula for each dimension *i* simultaneously either using the numerical integral solution or using the simulation process as we will demonstrate later. At first, we will try to write Eq. (5-6) in terms of copula, which allows us to solve the integral of the joint distribution and to seek for a non-linear and asymmetric pattern of relationships among the error terms which give more flexibility in modeling.

5.4.2 Copula with Count Data

Sklar's theorem (1959) states that there exists a class of distribution function that the n-dimensional cumulative distribution can be expressed in terms of the copula and the marginal. Copulas allow for a wide range of marginal distributions, it allows for the use of dependence using negative binomial models for overdispersed data, and as we mentioned previously, it allows for non-linear and asymmetric patterns of relationships among the marginals.

When y_{iq} are discrete (count) variables and F_i are discrete cdf's, the multivariate cumulative probability density function $H_{(1,2,...,I)}$ for a given segment *q* (as shown in Eq.(5-5)) can be constructed from $\left[F_1(y_{1q}),F_2(y_{2q}),....,F_I(y_{Iq})\right]\in Ran(F_1)\times Ran(F_2)\times....\times Ran(F_I)$, where *Ran*(F_i) denotes the 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:

$$\begin{aligned} &H_{(1,2,...,I)q}\left(\eta_1^{k_{1q}},\eta_2^{k_{2q}},....,\eta_I^{k_{Iq}}\right) \\ &= H_{(1,2,...,I)q}\left\{H_1^{-1}\left[\Pr\left(y_{1q}\le k_{1q}\right)\right],H_2^{-1}\left[\Pr\left(y_{2q}\le k_{2q}\right)\right],....,H_I^{-1}\left[\Pr\left(y_{Iq}\le k_{Iq}\right)\right]\right\} \\ &= C_{(1,2,...,I)q}\left[F_1\left(k_{1q}\right),F_2\left(k_{2q}\right),....,F_I\left(k_{Iq}\right)\mid \boldsymbol{\theta}\right] \end{aligned} \quad (5\text{-}7)$$

For all $\left[F_1(y_{1q}),F_2(y_{2q}),....,F_I(y_{Iq})\right]\in[0,1]^I$; $\boldsymbol{\theta}$ 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. (6) we can get:

$$\frac{\partial H_{(1,2,\ldots,I)q}\left(\varepsilon_{1q},\varepsilon_{2q},\ldots,\varepsilon_{Iq}\right)}{\partial\varepsilon_{1q},\partial\varepsilon_{2q},\ldots,\partial\varepsilon_{Iq}}=\frac{\partial C\left[F_1\left(y_{1q}\right),F_2\left(y_{2q}\right),\ldots,F_I\left(y_{Iq}\right)\right]}{\partial F_1\left(y_{1q}\right),\partial F_2\left(y_{2q}\right),\ldots,\partial F_I\left(y_{Iq}\right)}$$
$$f_{(1,2,\ldots,I)q}\left(\varepsilon_{1q},\varepsilon_{2q},\ldots,\varepsilon_{Iq}\mid\boldsymbol{\theta}\right)=c_{(1,2,\ldots,I)}\left[F_1\left(y_{1q}\right),F_2\left(y_{2q}\right),\ldots,F_I\left(y_{Iq}\right)\right].\prod_{i=1}^{I}f_i\left(y_{iq}\right) \quad (5\text{-}8)$$

where $c_{(1,2,\ldots,I)}$ 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 Eq. (5-8) into Eq. (5-6) to get:

$$\begin{aligned}&\Pr\left(y_{1q}=k_{1q},y_{2q}=k_{2q},\ldots,y_{Iq}=k_{Iq}\right)\\&=\int_{F_1\left(k_{1q}-1\right)=H_1\left(\eta_1^{k_{1q}-1}\right)}^{F_1\left(k_{1q}\right)}\int_{F_2\left(k_{2q}-1\right)}^{F_2\left(k_{2q}\right)}\ldots\int_{F_I\left(k_{Iq}-1\right)}^{F_I\left(k_{Iq}\right)}c_{(1,2,\ldots,I)}\left[F_1\left(y_{1q}\right),F_2\left(y_{2q}\right),\ldots\right.\\&\left.\ldots,F_I\left(y_{Iq}\right)\mid\boldsymbol{\theta}\right].\prod_{i=1}^{I}f_i\left(y_{iq}\right)dF_1\left(y_{1q}\right)dF_2\left(y_{2q}\right).\ldots.dF_I\left(y_{Iq}\right).\end{aligned} \quad (5\text{-}9)$$

Eq. (5-9) 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, which offers a non-closed form for Eq. (5-9) and the integral should be evaluated either numerically or by simulation. Second, the Archimedean copula, which offers the closed-form and by taking differences of the copula function *C* for the same boundaries of Eq. (5-9). The model estimation is carried out after specifying a suitable marginal distribution *F* for the count outcome and an appropriate copula *C*.

The copula constructed from such a formulation is not unique like its analogy resulting from continuous outcomes and continuous marginal distributions. 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 corresponding to the continuous one. More details regarding this issue are presented by Zimmer and Trividi (2006), Denuit and Lambert (2005) and Genest and Nešlehová (2007).

5.4.3 Marginal Distribution Selection

One of the main problems in modeling the count data usually is presence of greater variability (statistical overdispersion) in the count dataset. This problem

occurs when the observed variance is larger than the variance of a theoretical model. Utilizing copula concept permits to specify the marginal distribution in a more flexible way to accommodate this problem. For this purpose, the crash count type y_{iq} is assumed to follow a negative binomial type-II distribution (NBII). Each marginal for each crash type is determined conditionally on a vector of explanatory variables $\mathbf{x}_{iq}$ (not a mandatory to be a same set). $\mathbf{x}_{iq}$ affects a certain type of crash *i* with corresponding to set of parameter vectors $\boldsymbol{\beta}_i$. For each observation, the NBII *cdf* is obtained by summing the crash numbers from 0 to k_{iq} as:

$$\begin{aligned}\Pr\left(y_{iq} \le k_{iq} \mid \mathbf{x}_{iq}, \boldsymbol{\beta}_i\right) &= F_i\left(k_{iq}\right) = \sum_{r=0}^{k_{iq}} f_i\left(r \mid \mathbf{x}_{iq}, \boldsymbol{\beta}_i\right) \\ &= \sum_{r=0}^{k_{iq}} \left[\frac{\Gamma\left(r+\psi_i\right)}{\Gamma\left(\psi_i\right)\Gamma\left(r+1\right)} \left(\frac{\psi_i}{\lambda_{iq}+\psi_i} \right)^{\psi_i} \left(\frac{\lambda_{iq}}{\lambda_{iq}+\psi_i} \right)^{r} \right]\end{aligned} \tag{5-10}$$

where $\lambda_{iq} = \exp\left(\boldsymbol{\beta}_i' \mathbf{x}_{iq}\right)$; λ_{iq} is the conditional mean and $1/\psi_i$ is overdispersion parameter in the conditional variance $\Gamma_i\left(y_{iq}\right) = \lambda_{iq} + \psi_i^{-1}\left(\lambda_{iq}\right)^2$ (overdispersion occurs when $1/\psi_i > 0$).

5.4.4 Choosing a Copula Function

As previously stated, several types of parametric copula functions (Archimedean and Elliptical) are available for model development. The issue of choosing which best copula function for fitting data has been an attention-grabbing topic in the literature. So far, there is no robust formula that assess the goodness-of-fit of a copula without the need to investigate all the other types of copulas. This can be done through graphical techniques (four available) to assess our selection of the parametric copula based on rank of the observation and use a scatter plot concept. These techniques are not used for goodness-of-fit rather than they give an initial insight of the tendencies of the dependency structure of the outcomes regardless of the marginal effect. These techniques are: a) The PP-plot which is most general, least effective. b) Tail dependence plot – general and more effective than PP-plot. c) K-plot which is used for Archimedean copulas only and finally, d) t-plot which is most

restrictive - good 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 “Copula PP-plot” is introduced by Sun, Frees and Rosenberg (2008) inspired by the univariate PP-plot. 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:

$$LR(Z) = \begin{cases} L(Z) & \text{if} \quad 0 \le Z < 0.5 \\ R(Z) & \text{if} \quad 0.5 \le Z \le 1 \end{cases} \tag{5-11}$$

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

$$\begin{aligned} L(Z) &= \frac{\Pr\left(y_i \le F_i^{-1}(Z), y_j \le F_j^{-1}(Z)\right)}{\Pr\left(y_i \le F_i^{-1}(Z)\right)} = \frac{C(Z,Z)}{Z} \\ R(Z) &= \frac{\Pr\left(y_i > F_i^{-1}(Z), y_j > F_j^{-1}(Z)\right)}{\Pr\left(y_i > F_i^{-1}(Z)\right)} = \frac{C(1-Z,1-Z)}{1-Z} = \frac{1-2Z+C(Z,Z)}{1-Z} \end{aligned} \tag{5-12}$$

As implied above, we need to construct the empirical copula at first to conduct these graphical techniques. Let (m_{iq}, m_{jq}) a pair of observed crash count for type *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 $V_b \in [0,1]$ consequently, it can be formulated as:

$$\tilde{C}^{n}_{(i,j)}\left[\frac{a}{Q},\frac{b}{Q}\right]=\frac{\text{number of pairs }\left(m_{iq},m_{jq}\right)\text{ with }m_{iq}\le m_{i(a)},m_{jq}\le m_{j(b)}}{Q}$$
$$=\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{I}\left(\tilde{F}^{n}_{i}\left(m_{iq}\right)\le U_a,\tilde{F}^{n}_{j}\left(m_{jq}\right)\le V_b\right) \quad (5\text{-}13)$$
$$=\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{I}\left(R_i\left(m_{iq}\right)\le a,R_j\left(m_{jq}\right)\le b\right)$$

where $m_{i(a)}$ and $m_{j(b)},1\le a,b\le Q$ are order statistics from the sample, $\tilde{F}^{n}_{i}\left(m_{iq}\right)=\frac{1}{Q}\sum_{f=1}^{Q}\mathrm{I}\left(m_{if}\le m_{iq}\right)$ and $\tilde{F}^{n}_{j}\left(m_{jq}\right)=\frac{1}{Q}\sum_{f=1}^{Q}\mathrm{I}\left(m_{jf}\le m_{jq}\right)$ are the empirical cumulative distribution functions of the observations, $R_i\left(m_{iq}\right)$ and $R_j\left(m_{jq}\right)$ are the rank functions[2] 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. Types of available empirical copulas are shown in Table (5-2), in which 1/Q type is used in this study (see Asquith (2016) and Hernandez-Maldonado et al., (2012) for more details). The tail dependence of the empirical copula is constructed using Eq.(5-12).

Table (5-2) Types of the avaible empirical copulas

Type	Formulation
1/Q	$\tilde{C}^{n}_{(i,j)}\left[U_a,V_b\right]=\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{I}\left(\frac{R_i\left(m_{iq}\right)}{Q}\le U_a,\frac{R_j\left(m_{jq}\right)}{Q}\le V_b\right)$
Hazen	$C^{Ha}_{(i,j)}\left[U_a,V_b\right]=\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{I}\left(\frac{R_i\left(m_{iq}\right)-0.5}{Q}\le U_a,\frac{R_j\left(m_{jq}\right)-0.5}{Q}\le V_b\right)$
Weibull	$\tilde{C}^{W}_{(i,j)}\left[U_a,V_b\right]=\frac{1}{Q}\sum_{q=1}^{Q}\mathrm{I}\left(\frac{R_i\left(m_{iq}\right)}{1+Q}\le U_a,\frac{R_j\left(m_{jq}\right)}{1+Q}\le V_b\right)$
Bernstein	$\tilde{C}^{B}_{(i,j)}\left[U_a,V_b\right]=\sum_{a=1}^{Q}\sum_{b=1}^{Q}\tilde{C}^{n}_{(i,j)}\left[U_a,V_b\right]\times\eta\left(a,b;U_a,V_b\right)$, and $\eta\left(a,b;U_a,V_b\right)=\binom{Q}{a}\left[U_a\right]^{a}\left[1-U_b\right]^{Q-a}\times\binom{Q}{b}\left[V_b\right]^{b}\left[1-V_b\right]^{Q-b}$

[2] It is straight forward to see that the rank function is expressed as $R_i\left(m_{iq}\right)=\sum_{f=1}^{Q}\mathrm{I}\left(m_{if}\le m_{iq}\right)$, given that $\tilde{F}^{n}_{i}\left(m_{iq}\right)=R_i\left(m_{iq}\right)/Q$, therefore, the empirical copula can be seen as the empirical distribution of the rank transformed data as shown in the last part of Eq. (5-13)

5.4.5 Composite Marginal Likelihood CML

The CML approach has been used to overcome the multi-dimensionality of dependencies of Eq. (5-9) without a need to evaluate the full likelihood function. The CML approach works perfectly with copula due to the fact that there is a huge class of closed-form bivariate copula without a need to sacrifice any levels of dependency compared to other approaches. There are several methods, nested under the CML approach, for this type of model, we will use the pairwise marginal likelihood estimation method (see Castro et al., 2013; Castor et al., 2012; Bhat et al., 2014c; Yamamoto and Morikawa, 2013; Sener et al., 2010; Paleti et al., 2013 and Ferdous et al., 2010). The features of bivariate copula can be obtained from Eq. (5-7) when *I* = 2 with the following properties $C\left[F_1(y_{1q}),0\right] = C\left[0,F_2(y_{2q})\right] = 0$; $C\left[F_1(y_{1q}),1\right] = F_1(y_{1q})$ and $C\left[1,F_2(y_{2q})\right] = F_2(y_{2q})$. Let $(m_{1q}, m_{2q},....,m_{Iq})$ and $\zeta = (\beta_1,\beta_2,....,\beta_I;\psi_1,\psi_2,....,\psi_I;\boldsymbol{\theta})$ represent the actual observed crash count of type *i* on a specific segment *q* and a parameter vector of MCORC model, respectively. Let also $(j = 1,2,....,I)$, and Eq. (5-9) collapses into *I* (*I*-1)/2 pairs of bivariate probability computations and it takes the form:

$$
\begin{aligned}
L_{CML_q}(\zeta) &= \prod_{i=1}^{I-1}\prod_{j=i+1}^{I}\Pr\left(y_{iq}=m_{iq},y_{jq}=m_{jq}\right)\\
&= \prod_{i=1}^{I-1}\prod_{j=i+1}^{I}\left\{\int_{F_i(m_{iq}-1)}^{F_i(m_{iq})}\int_{F_j(m_{jq}-1)}^{F_j(m_{jq})} c_{(i,j)}\left[F_i(y_{iq}),F_j(y_{jq})\,|\,\theta_{ij}\right].\prod_{i=1}^{I} f_i(y_{iq})\,dF_i(y_{iq})dF_j(y_{jq})\right\} \quad (5\text{-}14)\\
&= \prod_{i=1}^{I-1}\prod_{j=i+1}^{I}\left\{\begin{matrix} C\left[F_i(m_{iq}),F_j(m_{jq})\,|\,\theta_{ij}\right]-C\left[F_i(m_{iq}-1),F_j(m_{jq})\,|\,\theta_{ij}\right] \\ -C\left[F_i(m_{iq}),F_j(m_{jq}-1)\,|\,\theta_{ij}\right]+C\left[F_i(m_{iq}-1),F_j(m_{jq}-1)\,|\,\theta_{ij}\right]\end{matrix}\right\}
\end{aligned}
$$

where θ_{ij} represents the level of dependency between the marginals $F_i(m_{iq}), F_j(m_{jq})$ for a certain copula function *C*. The pairwise marginal likelihood estimation across all segments can be computed using $L_{CML}(\zeta) = \prod_{q=1}^{Q} L_{CML_q}(\zeta)$.

The pairwise likelihood function above is easy to maximize, where the pairwise estimator $\hat{\zeta}_{CML}$ obtained by maximizing the logarithm of the $L_{CML}(\zeta)$ function with respect to the vector ζ which is consistent and asymptotically normal distributed with asymptotic mean ζ and covariance matrix given by the inverse

of Godambe's (1960) sandwich information matrix $\mathbf{G}(\zeta)$ (see Zhao and Joe, 2005; Castro et al., 2012 and Ferdous et al. 2010).

$$\begin{aligned}\mathbf{V}_{CML}(\zeta) &= [\mathbf{G}(\zeta)]^{-1} \\ &= [\mathbf{H}(\zeta)]^{-1}\mathbf{J}(\zeta)[\mathbf{H}(\zeta)]^{-1}, \text{where} \\ \mathbf{H}(\zeta) &= E\left[-\frac{\partial^2 \ln L_{CMLq}(\zeta)}{\partial\zeta\,\partial\zeta'}\right] \text{and} \\ \mathbf{J}(\zeta) &= E\left[\left(\frac{\partial \ln L_{CMLq}(\zeta)}{\partial\zeta}\right)\left(\frac{\partial \ln L_{CMLq}(\zeta)}{\partial\zeta'}\right)\right]\end{aligned} \tag{5-15}$$

where $\mathbf{H}(\zeta)$ and $\mathbf{J}(\zeta)$ are Hessian and Jacobean matrices, respectively. The Hessian and Jacobean matrices can be estimated in a straightforward manner at the CML estimate ($\hat{\zeta}_{CML}$):

$$\begin{aligned}\hat{\mathbf{H}}(\hat{\zeta}) &= \frac{1}{Q}\sum_{q=1}^{Q}\left[-\frac{\partial^2 \ln L_{CMLq}(\zeta)}{\partial\zeta\,\partial\zeta'}\right]_{\hat{\zeta}} \\ &= \left[\frac{1}{Q}\sum_{q=1}^{Q}\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}\frac{\partial^2 \ln \Pr(y_{iq}=m_{iq}, y_{jq}=m_{jq})}{\partial\zeta\,\partial\zeta'}\right]_{\hat{\zeta}}, \text{and} \\ \hat{\mathbf{J}}(\hat{\zeta}) &= \frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\frac{\partial \ln L_{CMLq}(\zeta)}{\partial\zeta}\right)\left(\frac{\partial \ln L_{CMLq}(\zeta)}{\partial\zeta'}\right)\right]_{\hat{\zeta}}.\end{aligned} \tag{5-16}$$

comply with the results of the empirical copula, the bivariate parametric copula in Eq. (5-14) will be the same function among all the pairs of crash types in our proposed model, but also, we will try to investigate other copula functions to get the best fit as we will see later.

5.4.6 Interdependence Interpretation

The bivariate copula function includes usually 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 and no point of tie them together. 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, most 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 $[-1,+1]$. Bouyé et al. (2000) stated, however, it is not the case when y_{iq} is discrete-count variable. Marshall (1996) and Tajar et al. (2001) elucidated 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 an 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.

5.4.7 Model Estimation Selection

There is no assent about a statistic criterion that selects the copula that provides the best fit to the data. 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 which model fits better, but doesn't tell which model is statistically significant when compared all competing models. Therefore, a non-nested likelihood ratio test is also used (more details on this test see Ben-Akiva and Lerman, 1985). The log likelihood function in Eq. (5-14) was coded in Gauss Aptech (1999) and the default BFGS algorithm provided by the maxlik

procedure in Gauss was used for maximizing the log-likelihood function.

5.4.8 Variance Covariance Structure of MCORC 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,

$$\mathbf{V}_{I,I}=\begin{bmatrix}\Gamma_1 & & & & \\ \Omega_{(2,1)} & \Gamma_2 & & & \\ \Omega_{(3,1)} & \Omega_{(3,2)} & \Gamma_3 & & \\ \vdots & \vdots & \vdots & \ddots & \\ \Omega_{(I,1)} & \Omega_{(I,2)} & \Omega_{(I,3)} & \Omega_{(I,I-1)} & \Gamma_I\end{bmatrix} \tag{5-17}$$

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 D'Angelo et al., 2013 and Hoeffding, 1940). Hoeffding's formula for the expected covariance between two continuous dependent variables x_i and x_j states that

$$\Omega\left(x_i,x_j\right)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left\{C\left[F_i\left(x_i\right),F_j\left(x_j\right)|\,\theta_{ij}\right]-\left[F_i\left(x_i\right)\times F_j\left(x_j\right)\right]\right\}dx_i dx_j \tag{5-18}$$

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

$$\begin{aligned}\Omega_{(i,j)}\left(y_{iq},y_{jq}\right)&=\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\Pr\left(y_{iq}\le r,y_{jq}\le s\right)-\left[\sum_{r=0}^{\infty}\Pr\left(y_{iq}\le r\right)\right]\times\left[\sum_{s=0}^{\infty}\Pr\left(y_{jq}\le s\right)\right]\\&=\left\{\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}C\left[F_i(r),F_j(s)|\,\theta_{ij}\right]-\left[\sum_{r=0}^{\infty}F_i(r)\right]\times\left[\sum_{s=0}^{\infty}F_j(s)\right]\right\}.\end{aligned} \tag{5-19}$$

More details on the above formula are given as a result of Eq. (B.1-7) in **Appendix.**B.) The average of the expected covariance $\Omega_{(i,j)}$ of all segments is calculated using,

$$E\left[\Omega_{(i,j)}\left(y_i, y_j\right)\right] = \frac{1}{Q}\sum_{q=1}^{Q}\Omega_{(i,j)}\left(y_{iq}, y_{jq}\right) \tag{5-20}$$

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

$$Cov\left(y_i, y_j\right) = Cov\left(\lambda_i, \lambda_j\right) + E\left[\tilde{\Omega}_{(i,j)}\left(y_i, y_j\right)\right] \tag{5-21}$$

In similar way, the variances element Γ_i in the diagonal variance-covariance matrix is calculated for each crash type *i* for the NBII model as, $\Gamma_i\left(y_{iq}\right) = \lambda_{iq} + \psi_i^{-1}\left(\lambda_{iq}\right)^2$ 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,

$$E\left[\Gamma_i\left(y_i\right)\right] = \frac{1}{Q}\sum_{q=1}^{Q}\Gamma_i\left(y_{iq}\right) \tag{5-22}$$

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

$$V_T\left(y_T\right) = V\left[E\left(y_T\right)\right] + E\left[V\left(y_T\right)\right] \tag{5-23}$$

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, both components are given by Eq. (5-24) and Eq. (5-25) respectively.

$$V\left[E\left(y_T\right)\right] = \sum_{i=1}^{I} Var\left(\lambda_i\right) + 2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} Cov\left(\lambda_i, \lambda_j\right) \tag{5-24}$$

$$E\left[V\left(y_T\right)\right] = \sum_{i=1}^{I} E\left[\Gamma_i\left(y_i\right)\right] + 2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I} E\left[\Omega_{(i,j)}\left(y_i, y_j\right)\right] \tag{5-25}$$

5.5 EMPIRICAL CRASH DATA SETTING

The crash data used in the analysis are used in same configuration to the previous chapter. The crash-count type distributions are presented in Figure (3-2) while the descriptive statistic of the main explanatory variables in this study is shown in Table (3-2).

5.6 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 MCORC model to our dataset. It is an interest point to investigate whether these types of crashes are jointly determined. The benefit of getting the joint distribution probability of these types together lies in the enhancement of the parameter estimates; furthermore, it helps to predict the number of crashes of each type simultaneously rather than individually. Followed by more investigation on the variance 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.

5.6.1 Empirical Copula Diagnosis

The empirical copula is formulated using Eq. (5-13) for each pair of the designated crash types. The empirical copula[3] of rear end and sideswipe crash types pair is given in Figure (5-1) (other pairs are reported in Appendix.A). We used the 1/Q empirical copula for our estimation and it was compared to a selected parametric copula (Frank) as an example (other types of empirical copula like Hazen; Weibull and Bernstein showed no much differences from our selected empirical copula for all other pairs). We investigated five 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 (5-2) shows the PP-plot for the same pair we used for the empirical copula and repeated for each parametric copula that we assigned to our developed model later. With several competing parametric copulas, we prefer the one that is closest to the empirical in some sense. We can see that Gaussian; Frank and Gumbel are a good starting for our modeling

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

for this pair. We repeated this plot for other pairs and all shows same conclusion (Other pairs are not reported in this paper). The strength of the PP-plots is available regardless of the dimensions (number of the count outcomes *l*) while the weakness can be seen as the lack to distinguish among different PP-plots due to the cumulative process can conceal some important differences in the distributions (Gibbons and Chakraborti, 2011). Therefore, we conducted the tail dependence to investigate the tendency of the upper/lower tails of the crash count type's distributions. The tail dependence plot depends only on the empirical copula and so it is not restricted to a specific class of copulae. The tail dependence of the rear end and the sideswipe crash types is shown in Figure (5-3) We can see that most of the observations given by the empirical copula are located in the upper tail (segments with low number of crash count) with a pattern almost similar to Frank; Gaussian copula. The same results can be deduced from other pairs (other pairs are reported in Appendix.A).

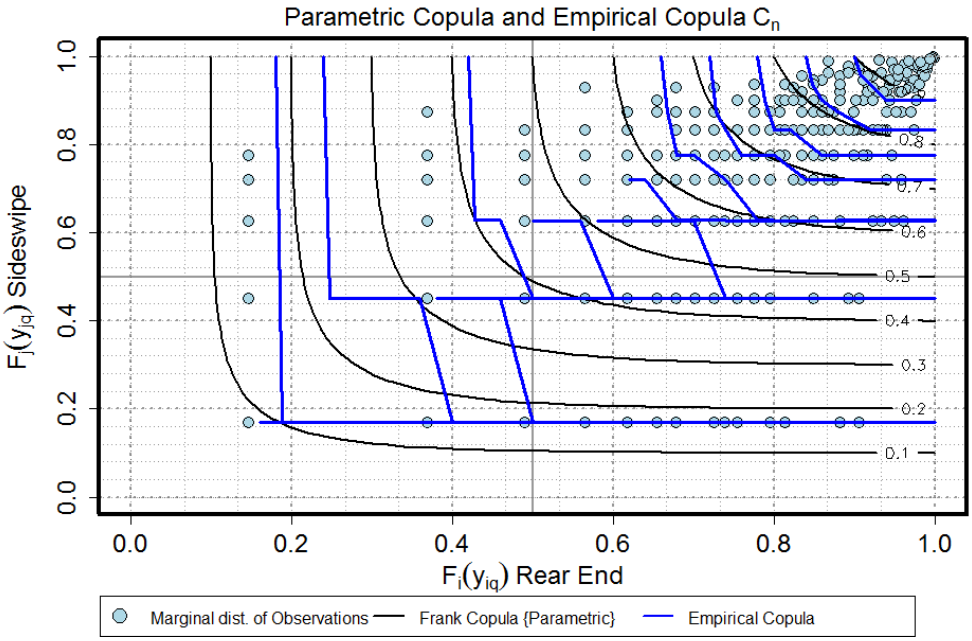


Figure (5-1) The empirical copula using 1/Q type compared to a selected parametric copula

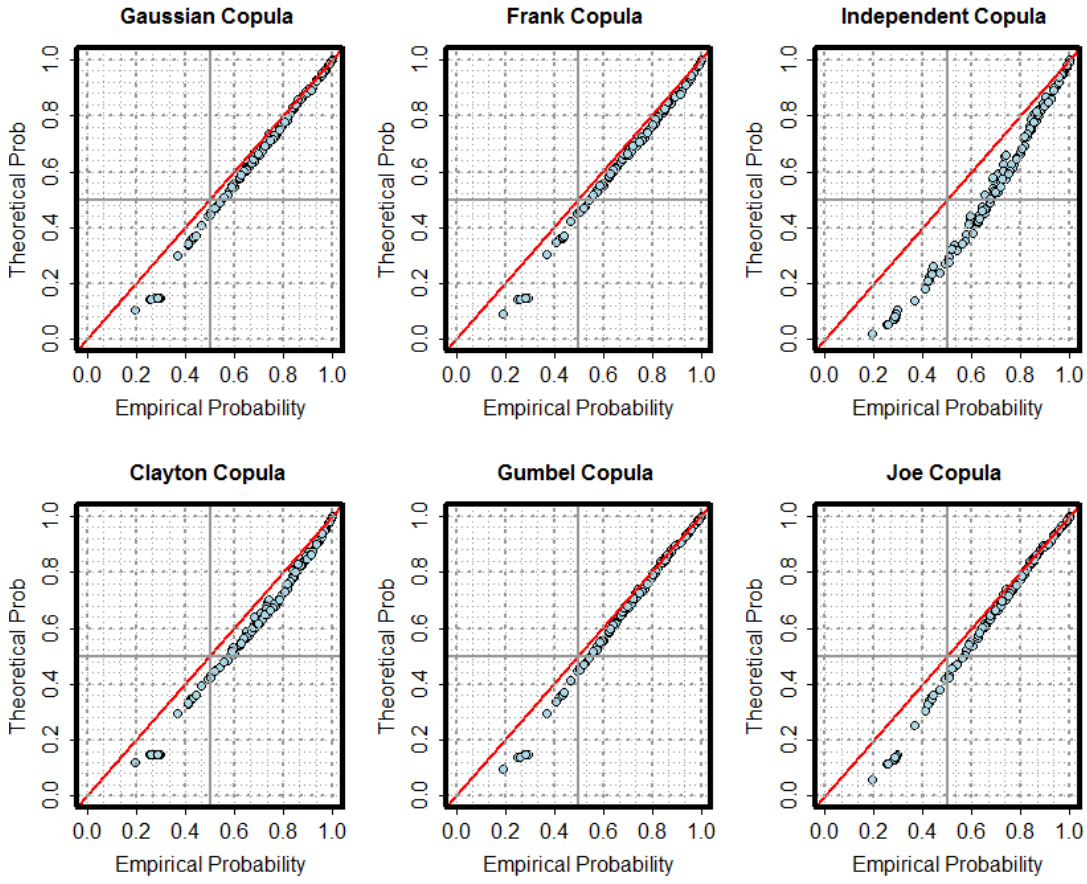


Figure (5-2) PP-plot of the parametric copula vs. the empirical copula.

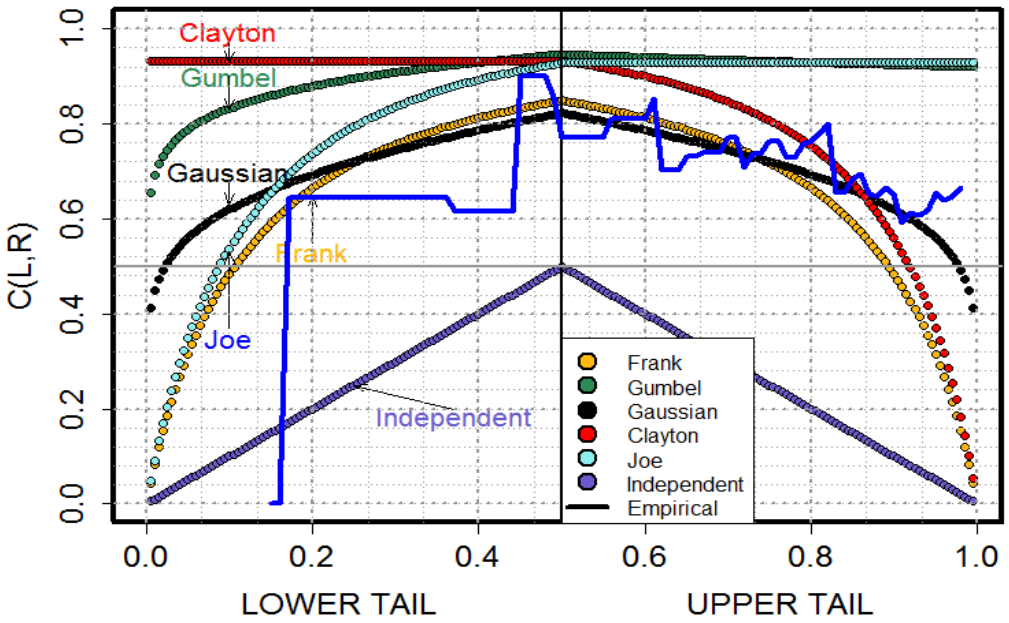


Figure (5-3) Tail dependence of the parametric copulas vs. the empirical copula

5.6.2 Model Specification and Crash Data Fitting

Let y_{iq} denote 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. The empirical model considers 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 $\boldsymbol{\beta}_i$. Regardless the conclusion we got from the empirical copula we used the same copula are implemented to model the unobserved heterogeneity ε_{iq} that generated from each crash type latent variable. The 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.

Identifying the most significant explanatory variables $\mathbf{x}_{iq}$ for each crash type is a prerequisite since each crash type has its distinct mechanisms and characteristics. For this purpose, the 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 copula functions. The preliminary estimation[4] of the independent copula suggests using NBII as a marginal for 'all-others' crash type category is not suitable since we got a non-significant dispersion parameter ψ_4 .The Poisson's marginal was selected instead for the 'all-others' type and it will be kept the same when

[4] 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. Gauss reference manual (Aptech, 1998) states that maximum allowed number of both gamma/factorial function argument, should be set no more 170. Obviously, this is not the case in our crash count dataset (e.g., in one of observed segments of the freeway, rear end crash is recorded to 212. avoiding the overflow problem if the logarithm of both gamma/factorial function is used. Thus, to facilitate the speed of computation, a Strling's formula offers an approximation in this regard (see Winkelmann, (2008) for more details).

investigating different types of copula function. The estimation results of the independent copula are presented in Table (5-3). It is worth to mention that the independent copula can be obtained [5] also by setting all the correlation components θ_{ij} of Gaussian copula to zero.

Gaussian; Frank; Clayton; Gumbel and Joe copula are implemented to fit our crash types data using MCORC model, but only the best copula in a statistical point of view will be reported to conserve on space. The copula function facilitates the correlation between the pairs of the marginal distributions of the crash types. The nested-likelihood ratio test is conducted between the Gaussian and the 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 $813.41\left(=2\times\left[-18271.76-\left(-17865.06\right)\right]\right)$, 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 model. Table (5-4) represents the performance measure of the log-likelihood; *AIC* and *BIC* of each copula function. It is clear that Frank 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 non-nested likelihood ratio test is also carried out to assess this conclusion by comparing the Frank-copula to the closest competitor Gaussian-copula model. The difference in the adjusted rho square ($\bar{\rho}_i^2$) value is 0.02608. The probability that this difference could have occurred by chance is less than $\Phi\left(-\left[-2\times 0.02608\times LL(C)+(38-38)\right]^{0.5}\right)$. This value, with $LL(C)=-45215.50$, is equal to $4.7578E-24$ which is almost zero, indicating that the difference in $\bar{\rho}_i^2$ between Frank and Gaussian models is statistically insignificant and that Frank copula is more suitable to fit our crash types count data.

[5] We can obtain the independent copula density function in a straightforward form by taking the difference of the copula between the two marginals F(y1) and F(y2) as c(F(y1), F(y2); θ) =[F(y1). F(y2)]- [F(y1-1). F(y2)]- [F(y1). F(y2-1)]- [F(y1-1). F(y2-1)],

Table (5-3) Multivariate Copula-Based Generalized Ordered Response Count Model of Crash Types: Independent Copula.

Explanatory Variables	Rear End		Sideswipe		Fixed Object		All-others	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Constant	-24.959	2.216	-16.505	1.220	-7.017	1.805	-11.040	1.021
LnAADT	2.651	0.230	1.723	0.127	0.800	0.173	1.170	0.108
LnLength	0.597	0.124	0.708	0.089	0.882	0.137	0.973	0.059
Urban rural dummy, 1 if rural, 0 if urban	-0.628	0.142	-0.710	0.134	-0.193	0.124‡		
Proportion of three or more lanes cross section by length of segment	0.937	0.099	0.606	0.111	0.363	0.090	0.418	0.102
Number of horizontal curves per segment	0.141	0.055	0.116	0.030	0.068	0.035†	-	-
Diamond interchange type dummy	-0.239†	0.142	-0.275	0.079	-0.239	0.104	-	-
Smallest vertical gradient in segment	-	-	0.073	0.046‡	-	-	-	-
Largest beginning vertical curve elevation in segment	-	-	0.878	0.744‡	-	-	-	-
Largest horizontal curve central angle in segment	-	-	-	-	0.507	0.269†	-	-
Number of vertical curves in segment	-	-	-	-	-0.047	0.036‡	-	-
□ (overdispersion parameter)	1.720	0.134	4.606	0.759	3.601	0.673	-	-
Sample size	822							
LL(C) with constant parameters only	-40,815.0							
LL(B) at convergence	-18,271.8							
Adjusted rho square $\bar{\rho}_i^2$	0.551							
AIC	36,607.5							
BIC	36,758.3							

S.E. standard error; - Not relevant; † Significant at 10% level; ‡ Significant at 15% level. All the other coefficients are significant at the level of 5%. Significance of the actual overdispersion parameter (1/ ψ) is estimated using the delta method. Significance of the overdispersion effect is very strong for all crash types, at or better than the 99.5% level.

Table (5-4) Log-Likelihood, Akaike Information Criterion and Bayesian Information Criterion for Various Copulas.

Copula Type	Log-Likelihood	AIC	BIC
Independent	-18,272	36,608	36,758
Gaussian	-17,865	35,805	35,986
Clayton	-17,871	35,817	35,996
Frank	-16,686	33,447	33,626
Gumbel	-17,882	35,840	36,019
Joe	-17,894	35,865	36,044

One of the possible reasons is that among all the features of bivariate Archimedean's copula functions that have been used in empirical application, only Frank copula is a comprehensive one and support the range $\theta_{ij} : \in [-\infty, +\infty]$.

A comprehensive bivariate copula allows both the positive and negative interdependence structure among the pairs of marginal distributions. Frank copula function shares the same feature with Gaussian copula by offering a symmetrical dependency with near linear in the center, but flattens more in the tails compared to the Gaussian, no wonder why is the latter was the nearest competitor (see Winkelmann, 2012).

5.6.3 Estimation Results

Estimation results of MCORC-Frank regression model are presented in Table (5-5) while the interdependence parameters to explain the correlation among the marginal distributions of crash types 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 parameter is estimated to be 1.827; 4.848 and 2.038 for rear end; sideswipe and other types respectively, which imply an overdispersion magnitude of 0.547; 0.206 and 0.490. These parameters are significantly significant as evidence by the large t-values. The MCORC-Frank model supports different sizes of overdispersion in order to represent the correlations of the unobserved heterogeneity for each frequency of crash type.

Table (5-5) Multivariate Copula-Based Generalized Ordered Response Count Model of Crash Types: Frank Copula.

Explanatory Variables	Rear End		Sideswipe		Fixed Object		All-others	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Constant	-24.589	1.845	-16.408	1.223	-8.223	2.183	-10.477	1.075
LnAADT	2.614	0.187	1.712	0.127	0.896	0.217	1.108	0.112
LnLength	0.576	0.101	0.697	0.099	0.904	0.101	0.951	0.067
Urban rural dummy, 1 if rural, 0 if urban	-0.623	0.155	-0.686	0.136	-0.168‡	0.171		
Proportion of three or more lanes cross section by length of segment	0.925	0.103	0.618	0.114	0.360	0.112	0.438	0.106
Number of horizontal curves per segment	0.132	0.034	0.106	0.030	0.067	0.031	-	-
Diamond interchange type dummy	-0.211†	0.109	-0.266	0.077	-0.214	0.091	-	-
Smallest vertical gradient in segment			0.071	0.027	-	-	-	-
Largest beginning vertical curve elevation in segment	-	-	0.733	0.344	-	-	-	-
Largest horizontal curve central angle in segment	-	-	-	-	0.490	0.168	-	-
Number of vertical curves in segment	-	-	-	-	-0.047†	0.028	-	-
□	1.827	0.145	4.848	0.748	2.038	0.303	-	-
Level of Dependency θij								
Sideswipe	2.323	0.341						
Fixed Object	0.782‡	0.485	1.091	0.244				
All-others	0.609	0.274	0.975	0.243	0.985	0.197		
Sample size	822							
LL(C) with constant parameters only	-45,215.5							
LL(B) at convergence	-16,685.7							
Adjusted rho square $\bar{\rho}_i^2$	0.630							
AIC	33,447.3							
BIC	33,626.4							

- Not relevant; † Significant at 10% level; ‡ Significant at 15% level. All the other coefficients are significant at the level of 5%.

The MCORC-Frank copula of type specific expected crash frequency functions utilized almost same set of explanatory variables for each crash type. The results of the proposed model suggest that there are obvious 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 both 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 layout of the horizontal curve variables in general plays an influence factor to these crash types, this influence part ranged from speed differentials and their associations with rear end to lane offsetting and sideswipe, and potential loss of control and roadside encroachments resulting in fixed object crash type as we saw in Chapter Three. The location of the observed segment represented by the rural/urban dichotomous variable has a negative coefficient for both the rear end and sideswipe. The results are intuitive as expected from these two types of crashes that both are not as high as they would be in urban contexts due to for example the congestion related effects. The footprint of the diamond intersection shape provides usually a wide space for adequate merge and weave distances, thus reducing the probability of rear end; sideswipe or fixed objects crashes, no wonder we got a negative coefficient for this variable. The vertical curvature characteristics are intermittently significant represented by - minimum vertical grade variable being significant in the sideswipe function, while the maximum vertical curve elevation variable is also significant for same crash type.

The smallest vertical curve gradient in segment is associated positively with sideswipe crash type (0.071) at 1% significance level. Another variable associated to sideswipe crash type is the largest beginning vertical curve elevation in segment. This variable (0.733) is found to be statistically significant at 1% level. Similar to the above-mentioned, the effect of curve gradient, the likelihood of sideswipe crashes can go up as the maximum curve elevation in segment increases. The horizontal curvature characteristics are represented by the largest horizontal curve central angle in segment (0.490) which is

significant at 1% level for fixed object crash type only. Miss-judging a horizontal curve by the drive is explained by this variable which is associated with vehicle run-off-road and cause the fixed object crash type usually

5.6.4 Representativeness of Variance and Covariance Structure

As mentioned earlier, the copula function in the MCORC model is used to tie the pairs of the marginal distributions of the crash types by permitting the error term to correlate for each pair. Each θ_{ij} represents a level of dependency which demonstrates the presence of common unobserved factors in the latent variable functions. Parameter estimates θ_{ij} of the MCORC-Frank copula model among type of crashes are presented in the end of Table (5-5). A significant positive value of θ_{ij} indicates an association between the unobserved factors of each crash type in the corresponding pair. Rear end vs sideswipe; rear-end vs all-others; sideswipe vs fixed objects; sideswipe vs all- others and fixed object vs all-others are found to be statistically significant, except the rear end vs fixed object pair which is found to be not significant, indicates no correlation between these two crash types. The non-parametric Kendall's 'τ' measure was utilized to interpret the level of dependency θ_{ij}, and the results are presented in Table (5-6).

Table (5-6) Parameter estimates of τ_{ij} of multivariate copula-based generalized ordered response count model of crash types: Frank copula model

	Rear-End	Sideswipe	Fixed Object
Sideswipe	0.247†		
Fixed Object	0.061	0.131	
Other Types	0.068	0.108	0.125

† Using the Frank's τ_{ij} transformation formula, $\tau_{ij} = 1 - \frac{4}{\theta_{ij}}\left[1 - D_1(\theta_{ij})\right]$ where $D_k(\theta_{ij})$ denotes to the Debye function given as $\frac{k}{\theta_{ij}^k}\int_0^{\theta_{ij}} \frac{t^k}{(e^t - 1)} dt, k = 1,2.$

The correlation ranges between 0.06 and 0.25, which demonstrates the presence of common unobserved factors association of the unobserved latent propensity for each crash type. These values are considered to be quite small due to the average numbers of each crash type are small as well.

The $\mathbf{V}_{I,I}$ matrix was calculated[6] for MCORC Frank copula considering the average values among all segments using both Eq. (5-19) and Eq.(5-22) it's equal to,

$$\mathbf{V}_{4,4} = \begin{bmatrix} 191.68 & & & \\ 64.84 & 7.41 & & \\ 22.43 & 5.96 & 5.15 & \\ 18.90 & 4.72 & 2.32 & 1.71 \end{bmatrix} \quad (5\text{-}26)$$

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. (5-19). The total covariance of crash types is calculated for MCORC-Frank and independent copula models using Eq.(5-21) and presented in Table (5-7). The results suggest that Frank copula represents more accurately the covariance structure among the crash types.

Table (5-7) Comparative Total Covariance from Frank and Independent Copulas.

Pair of crash types	Observed	Frank	Independent
Rear end and sideswipe	104.097	124.123	57.925
Rear end and fixed object	46.118	46.081	27.481
Rear end and other types	29.177	38.938	19.487
Sideswipe and fixed objects	13.730	12.130	7.552
Sideswipe and other types	8.885	9.883	5.272
Fixed objects and other types	5.304	4.843	3.090

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.(5-24) and Eq.(5-25) respectively. Total variance

[6] It is worth to mention that evaluating the expected covariance elements in Eq. (5-19) can be also done by Eq. (B.1-1), but the amount of calculation time increases as the maximum number of crashes for a certain type of crash and a given segment increases. This is because Eq. (5-19) requires to calculate the probability using the differences between the upper and lower bounds for each crash type pair as given in Eq. (5-14). 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.

components of both Frank and independent copulas are presented in Table (5-8. The results suggest that Frank copula also represents the total variance structure more accurately compared to the independent copula. Finally, we have calculated the estimated correlation for a given segment using the following formula (Ophem, 1999).

$$\rho\left(y_{iq}, y_{jq}\right) = \frac{\Omega\left(y_{iq}, y_{jq}\right)}{\sigma_{y_{iq}}, \sigma_{y_{jq}}} \tag{5-27}$$

Table (5-8) Variance Structure of Total Crashes.

	Observed Crashes	Frank	Independent
$V[E(y_T)]$		501.327	494.60
$E[V(y_T)]$		444.383	204.64
$V_T(y_T)$	881.92	945.711	699.24

The results are shown in Table (5-9. It's an interesting to compare the correlations value to our previous work (See Mothafer et al., 2016). It seems that the correlation values that produced by the MCORC-Frank and the one by the MVPGM (Multivariate Poisson Gamma Mixture) model are slightly same, expect for the heavily tailed relationship described by the rear end and the sideswipe pair. It's the limitation of the MVPGM model since the correlation parameter is equally distributed among the pairs, this limitation was overcome by introducing the copula here.

Table (5-9) Estimated correlation matrix for a given segment

	Rear-End	Sideswipe	Fixed Object
Sideswipe	0.555 (64.84†)		
Fixed Object	0.330 (22.43)	0.382 (5.96)	
Other Types	0.314 (18.90)	0.342 (4.72)	0.289 (2.32)

† Covariance of the error term value are between parentheses

5.6.5 Marginal Effects

The marginal effect values of all the explanatory variables of MCROC-Frank copula model for each crash type are presented in Table (5-10). It is conspicuous to see the marginal effect values of rear end crash type are larger in the 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 Chapter 4. Another fact we found is that the characteristics of the horizontal curves on the designated freeway segment increase the likelihood of these types of crashes occurring, when compared to the characteristics of vertical curves.

Table (5-10) Marginal Effects of Multivariate Copula-Based Generalized Ordered Response Count Model: Frank Copula

Explanatory variable	Rear-End	Sideswipe	Fixed Object	Other Types
LnAADT	22.62	4.64	1.66	1.85
LnLength	4.98	1.89	1.68	1.58
Urban rural dummy, 1 if rural, 0 if urban	-0.86	-0.99	-0.18	
Proportion of three or more lanes cross section by length of segment	8.00	1.67	0.67	0.73
Number of horizontal curves per segment	1.14	0.29	0.12	-
Diamond interchange type dummy	-1.82	-0.72	-0.40	-
Smallest vertical gradient in segment	-	0.19	-	-
Largest beginning vertical curve elevation in segment	-	1.98	-	-
Largest horizontal curve central angle in segment	-	-	0.91	-
Number of vertical curves in segment	-		-0.09	-

5.7 SUMMARY

This chapter presents a multivariate copula-based ordered response model for non-negative integer counts outcomes. The points of interest in the multivariate modeling are to investigate whether these outcomes are jointly determined and to enhance the parameter estimate efficiency. The advantages of the proposed model for count outcomes lie in: first, the model offers a joint distribution without any restrictions to accept both positive and negative correlations among the error term structure components. Second, no need for a simulation-based technique which is usually a computationally burdensome in the multivariate econometric model. 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 tie two count marginal distributions that reflect two different count outcomes. The proposed model is parametric; straightforward to implement and more flexible to allow parameterizing the count marginal distribution to reflect the effect of the explanatory variables that affect each count outcome. The proposed model also offers a non-linear asymmetric interdependence structure among error term components. The correlations among the error components is obtained from transferring the level of dependency of the copula function into a non-parametric Kendall's 'τ' measure.

The model framework is demonstrated for an empirical application to study four different categories of crash types that commonly occur on freeway segments located on highway No. 5 in the State of Washington. The proposed model is used to investigate the dependence structure among these categories of crash types which are a rear end; sideswipe; fixed object and 'all-other'. The aim is to get a better understanding on the nature of each crash type and its influence in order to adjust the safety policy and to enhance the parameter estimates of the explanatory variables. The effects of geometry and traffic characteristics of the freeway segments of each type of crash have been

investigated. We examined five different copula functions and selected the NBII as the marginal distribution to capture different sizes of the overdispersion, which is considered a common problem in the crash count distributions. The empirical results show that Frank copula is best to fit our data in a statistical point of view. In addition, considering the correlation among the unobserved heterogeneity is highly recommended to enhance the covariance and the variance structure estimation when they are compared to the corresponding observed ones. The marginal effect calculations give an insight that the characteristics of horizontal curves of the selected freeway increase the likelihood of rear end; sideswipe; fixed objects and all-other types of crashes compared to the characteristics of vertical curves.

The severity level is not considered in this chapter. One might think of utilizing another crash count dataset that contains the number of crashes by both crash type and severity level which offers richer insights into the differential impacts of various explanatory variables on the crashes. Furthermore, it would be also a significant work to investigate the multivariate serial correlation among the analysis period of a pooled crash count dataset.

Chapter 6

Random Effect Poisson Gamma Mixture Model

6.1 INTRODUCTION

This chapter presents a negative binomial crash sum model as an alternative for modeling time invariant heterogeneity in short crash data panels. Time invariant heterogeneity arising through multiple years of observation (2005-2007) for each segment is viewed as a common unobserved effect at the segment level, and typically treated with panel models involving fixed or random effects. Random effects model unobserved heterogeneity through the error term, typically following a gamma or normal distribution. We take advantage of the fact that gamma heterogeneity in a multi-period Poisson count modeling framework is equivalent to a negative binomial distribution for a dependent variable which is the summation of crashes across years. The Poisson panel model referred to in this paper is the random effects Poisson gamma (REPG).

6.2 BACKGROUND

In the classical approach of modeling the crash count data, it is customary to assume the data is cross-sectional in nature (see for example, Lord, 2000; Ivan et al., 2000; Lyon et al., 2003; Miaou and Lord, 2003; Lord et al., 2005; Miaou and Song, 2005). Alternatively, when time effects are considered, we have a combination of cross-sectional data and time series data (also known as the panel data) in which the duration of observations is included (Law et al., 2009; Kumara and Chin, 2004; Chin and Quddus, 2003 and Quddus, 2008). Time

effects shared across observations result in the efficiency problem (Shankar et al., 1995; Greene, 2003). Ignoring time invariant heterogeneity results in the consequence that parameters that are insignificant in reality will be incorrectly identified as significant, and therefore included in the model. A recent study by Mannering et al (2016) discusses the various form of unobserved heterogeneity and consequences for parameter estimation. Time invariant heterogeneity is identified as an important aspect of unobserved heterogeneity, and the study emphasizes the importance of further study on this subject.

Time invariant heterogeneity in crash data can be modeled by accounting for repeated observation effects through a negative multinomial density (Ulfarsson and Shankar, 2003). In this approach, the negative binomial density is modified to account for contributions from each time period. In the second approach, the error term across time observations is treated as a random effect (Shankar et al, 1998) that follows an arbitrary continuous distribution[7]. A more general approach is to treat the time effect via the year indicator as a random parameter in a random parameter count model. If some or all of the year indicators are random, then, it implies that the serial correlation effect is stochastic and not constant across years. The random effects model is a constrained version of this model, where the intercept alone is random. Sittikariya et al. (2005) proposed a zero inflated Poisson (ZIP) model to account for excess zeroes in the crash data. In their method they used the negative multinomial approach to adjust the standard errors. By comparing the negative multinomial standard errors and cross sectional negative binomial standard errors, they used a loading factor principle which represented the level of inflation in standard errors of the parameters estimates due to serial correlation. Neither of the approaches or other published literature on serial correlation in count models addresses the impact of cumulative exposure. Under cumulative exposure, one can visualize the crash model to be composed of multiple years of observation, as opposed to the conventional one-year window. With multiple

[7] An alternative approach instead uses an arbitrary discrete distribution representation of unobserved heterogeneity, which generates a class of models called finite mixture models (Cameron and Trivedi, 2001).

years of observation, the problem of excess zeros is potentially mitigated, while the problem of time series effects is now treated as a single cumulative effect as opposed to a common unobserved effect across time periods for any given segment.

The aim of this chapter is to therefore evaluate the parameters of a count model under the influence of time invariant heterogeneity and compare the magnitudes and standard errors with those from a cumulative crash count model. The cumulative crash count model, also termed here in this paper as the crash sum model would represent a single cross sectional analysis of crashes summed up across the entire time period.

6.3 EMPIRICAL CRASH DATA SETTING

We used same configuration of the crash types of both chapter four and chapter five. The data here are organized as a panel count data. The time plays a vital role to add more information to our proposed model. The unobserved heterogeneity now is considered in a dynamic status. The crash-count type distributions are presented in Figure (3-2) while the descriptive statistic of the main explanatory variables in this study is shown in Table (3-2).

6.4 METHODOLOGY

Modeling panel crash count data with random effects is usually started by assuming the error term α follows an arbitrary continuous distribution $f(\alpha)$ (for example gamma, or Gaussian distribution, Cameron and Trivedi, 2013). Let *i* be an index of crash types, *t* be an index representing the year of observation in the panel, and *q* be an index of segment number, respectively. Then we can write the joint probability of the observed crash count variable y_{itq} for a given crash type *i* observed during time *t* on segment *q* as:

$$\begin{aligned} P\left[y_{i1q}, y_{i2q}, \ldots, y_{iTq}\right] &= \int_0^\infty P\left[y_{i1q}, y_{i2q}, \ldots, y_{iTq} \mid \alpha_{iq}\right] f\left(\alpha_{iq}\right) d\alpha_{iq} \\ &= \int_0^\infty \left[\prod_{t=1}^{T} P\left[y_{itq} \mid \alpha_{iq}\right]\right] f\left(\alpha_{iq}\right) d\alpha_{iq} \end{aligned} \tag{6-1}$$

where α_{iq} represents an error term that is invariant with time for a specific crash

type *i* and segment *q* (also known as the multiplicative segment-specific effect). Let's assume the observed number of crashes k_{itq} is drawn from a Poisson distribution as,

$$\begin{aligned} &\left(y_{itq}=k_{itq}\right)\sim P\left[\eta_{itq}=\alpha_{iq}\times\exp\left(\mathbf{x}'_{itq}\boldsymbol{\beta}_i\right)\right], \\ &P\left[y_{itq}=k_{itq}\right]=\frac{e^{-\eta_{itq}}\times\eta_{itq}^{kitq}}{k_{itq}!} \end{aligned} \tag{6-2}$$

Where η_{itq} represents the expected number of crashes, $\mathbf{x}'_{itq}$ is a vector of explanatory variables that affect the number of crashes and $\boldsymbol{\beta}_i$ is a vector of parameters to be estimated including the constant. Let $\lambda_{itq}=\exp\left(\mathbf{x}'_{itq}\boldsymbol{\beta}_i\right)$, and α_{iq} follows a gamma distribution with mean equal to one and variance equal to $1/\gamma_i$. Then, we can write the likelihood function based on Eq. (6-1) as:

$$\begin{aligned} L\left[\boldsymbol{\beta}_i,\gamma_i\right] &= \prod_{q=1}^{Q}\left\{\frac{\gamma_i^{\gamma_i}}{\Gamma\left(\gamma_i\right)}\left(\prod_{t=1}^{T}\frac{\lambda_{itq}^{k_{itq}}}{k_{itq}!}\right)\int_0^{\infty}\exp\left(-\alpha_{iq}\sum_{t=1}^{T}\lambda_{itq}\right)\alpha_{iq}^{\sum_{t=1}^{T}k_{itq}}\alpha_{iq}^{\gamma_i-1}\exp\left(-\gamma_i\alpha_{iq}\right)d\alpha_{iq}\right\} \\ &= \prod_{q=1}^{Q}\left\{\frac{\gamma_i^{\gamma_i}}{\Gamma\left(\gamma_i\right)}\left(\prod_{t=1}^{T}\frac{\lambda_{itq}^{k_{itq}}}{k_{itq}!}\right)\int_0^{\infty}\exp\left(-\alpha_{iq}\left[\gamma_i+\sum_{t=1}^{T}\lambda_{itq}\right]\right)\alpha_{iq}^{\gamma_i+\sum_{t=1}^{T}k_{itq}-1}d\alpha_{iq}\right\} \\ &= \prod_{q=1}^{Q}\left\{\left(\prod_{t=1}^{T}\frac{\lambda_{itq}^{k_{itq}}}{k_{itq}!}\right)\frac{\Gamma\left(\gamma_i+\sum_{t=1}^{T}k_{itq}\right)}{\Gamma\left(\gamma_i\right)}\left(\frac{\gamma_i}{\gamma_i+\sum_{t=1}^{T}\lambda_{itq}}\right)^{\gamma_i}\left(\frac{1}{\gamma_i+\sum_{t=1}^{T}\lambda_{itq}}\right)^{\sum_{t=1}^{T}k_{itq}}\right\} \end{aligned} \tag{6-3}$$

Letting $\omega_{iq}=\gamma_i\Big/\left(\gamma_i+\sum_{t=1}^{T}\lambda_{itq}\right)$, then the log-likelihood function for a crash type *i* and across all segments would be:

$$\begin{aligned} LL\left[\boldsymbol{\beta}_i,\gamma_i\right] = \sum_{q=1}^{Q}\Bigg\{&\log\Gamma\left(\gamma_i+\sum_{t=1}^{T}k_{itq}\right)-\sum_{t=1}^{T}\log\Gamma\left(1+k_{itq}\right)-\log\Gamma\left(\gamma_i\right)+\gamma_i\log\omega_{iq} \\ &+\log\left(1-\omega_{iq}\right)\sum_{t=1}^{T}k_{itq}+\sum_{t=1}^{T}k_{itq}\log\lambda_{itq}-\left(\sum_{t=1}^{T}k_{itq}\right)\log\left(\sum_{t=1}^{T}\lambda_{itq}\right)\Bigg\} \end{aligned} \tag{6-4}$$

Eq. (6-3) is the REPG model which is proposed by Hausman et al (1984). The

REPG distribution has $E(y_{itq}) = \lambda_{itq}$ and $V(y_{itq}) = \lambda_{itq} + \lambda_{itq}^2 / \gamma_i$ properties, and is equivalent to an NB distribution for $K_{iq} = \sum_{t=1}^{T} k_{itq}$ with expectation $\Lambda_{iq} = \sum_{t=1}^{T} \lambda_{itq}$. The log-likelihood function represented by Eq. (6-4) is coded in STATA (2005) and the default BFGS algorithm provided by the maxlik procedure in STATA was used for maximizing this function.

In the event that the **X** vector has few and infrequent and minor changes in value over T, assuming equality of **X** over the individual time periods, we see that

$$\ln \Lambda_{iq} = \ln \sum_{t=1}^{T} \lambda_{itq} = \ln(T\lambda_{itq}) \tag{6-5}$$

which implies that for a crash sum model, using a representative $\mathbf{x}_{iq}$, we get

$$\ln \Lambda_{iq} = \boldsymbol{\beta}'_{CS}\mathbf{x}_{iq} = \ln(T) + \boldsymbol{\beta}'_{REPG}\mathbf{x}_{itq} \tag{6-6}$$

where $\boldsymbol{\beta}'_{CS}$ is the parameter vector for the crash sum model, and $\boldsymbol{\beta}'_{REPG}$ is the estimated parameter vector for the REPG model, $\mathbf{x}_{iq}$ and $\mathbf{x}_{itq}$ are respective vectors of independent variables for the crash sum and REPG models. This relationship indicates that the crash sum model can be estimated with an adjustment to the constant by a scalar value of ln(T) to make parameter estimation consistent with the REPG. This is a reasonable adjustment in short panels, while in longer panels the applicability may not be suitable due to the fact that the modifications to the **X** vector can be significant (due to lane addition or width adjustments, ADT adjustments, shoulder width adjustments and curvature adjustments).

We discuss below results from the panel REPG model and the crash sum model. To begin the evaluation, the REPG is first baselined against a cross-sectional NB model (sample size N=822) denoted as the independent model, because this model does not assume sharing of time invariant heterogeneity across multiple time periods for a segment. In other words, the within-segment correlation is ignored and within-segment observations are

treated as independent observations. The idea is to illustrate the downward bias in standard error when time invariant heterogeneity is ignored. The apparent serial correlation induced by the time invariant heterogeneity can be substantial, as the discussion of the results will indicate. To keep the panel duration short, we select a three-year panel that experienced minimum changes in the **X** vector. We then evaluate the cumulative crash count model (sample size N=274). We estimate five crash outcome types in this evaluation, including total crashes, rear ends, sideswipes, fixed objects and all-other types. Therefore, in total, in this evaluation, five crash type models were estimated separately for: the independence (no serial correlation) scenario; REPG model scenario; and the cumulative crash count scenario.

6.5 RESULTS

The evaluation of the REPG model against the independent NB model is presented first, as shown in Table (6-1). The REPG model outperforms the independent NB model on several information criteria such as the AIC and BIC (Kuha, 2004). Most noticeably, the standard errors of the parameters in the independent model are significantly downward biased, with the exception of the overdispersion parameter. The overdispersion parameter spuriously captures the time invariant heterogeneity effect with an inflated standard error. The parameter estimates including the constant are downward biased by 40-60%, with horizontal curve count recording the highest bias of around 60%. This indicates a significant amount of time invariant heterogeneity.

The REPG model by assumption however assumes the explanatory effects are identical and consistent across the multiple time periods. The second issue relates to the empirical equivalence of the crash sum model and the REPG model. To evaluate the first issue relating to parameter constancy across multiple time periods, we estimated one year NB models for each of the analysis year, 2005, 2006 and 2007. We compared the parameter estimates year by year for these NB models and the asymptotic t-tests indicated that the similarity in parameters cannot be rejected at the 5% level of confidence[8]. The

[8] We took the chance to investigate this assumption even further by allowing the interaction of

second issue relating to the empirical equivalence of the crash sum model and the REPG was evaluated using the Hausman test (more details on Hausman's test see Baltagi, 2008; Cameron and Trivedi, 2005),

$$H = \left(\hat{\boldsymbol{\beta}}'^{REPG}_{Total} - \tilde{\boldsymbol{\beta}}'^{CS}_{Total}\right)' \left(\hat{V}\left[\hat{\boldsymbol{\beta}}'^{REPG}_{Total}\right] - \tilde{V}\left[\tilde{\boldsymbol{\beta}}'^{CS}_{Total}\right]\right)^{-1} \left(\hat{\boldsymbol{\beta}}'^{REPG}_{Total} - \tilde{\boldsymbol{\beta}}'^{CS}_{Total}\right) \quad (6\text{-}7)$$

This evaluation was done for all crash outcome models, but by way of example, we discuss the results for the total crash model, as shown in Eq.(6-7). The Hausman test here compares two models, the first being the crash sum model, and the second being the REPG model. The crash sum model under the null hypothesis is efficient, while the REPG model is consistent under the alternative. The parameter vectors $\hat{\boldsymbol{\beta}}'^{REPG}_{Total}$ $\tilde{\boldsymbol{\beta}}'^{CS}_{Total}$ denote the REPG model for total crashes and crash sum model for total crashes respectively. The middle term in Eq.(6-7) denotes the inverse of the variance difference for the two models. The results of Hausman's test (often denoted by *H*) produced an *H* statistic value of 4.63 which is not significant to reject the null hypothesis under 5% significance level of the critical value of chi-square distribution. This indicates that the null hypothesis cannot be rejected that the two models are equivalent and that the REPG model is a consistent alternative. Further, the asymptotic t-test comparison of parameters between the crash sum model and the REPG model indicated that no parameter was statistically different at the 5% level of confidence.

dummy variables of the years of observation along with the selected explanatory variables for each crash type and the total in the REPG model. The results show that the variability of the effects of the explanatory variables are not significant and the assumption is valid for our model purpose.

Table (6-1) Standard error downward bias in independent model of total crashes

Explanatory variables	REPG model		Independent NB model		Crash sum model	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
LnAADT	1.079**	0.122	1.542**	0.086	1.350**	0.124
LnLength	0.660**	0.088	0.754**	0.058	0.678**	0.086
Urban rural dummy, 1 if rural, 0 if urban	-0.463**	0.114	-0.323**	0.078	-0.340**	0.112
Proportion of three or more lanes cross section by length of segment	0.806**	0.092	0.749**	0.061	0.734**	0.089
Number of horizontal curves per segment	0.129**	0.032	0.104**	0.020	0.120**	0.031
Diamond interchange type dummy	-0.379**	0.082	-0.285**	0.054	-0.326**	0.081
Smallest vertical curve gradient in segment	0.117**	0.045	0.113**	0.029	0.112**	0.044
Shortest vertical curve length in segment in miles	-2.272**	0.839	-2.128**	0.577	-2.192**	0.819
Largest vertical curve rate of vertical in segment	-0.036†	0.019	-0.028*	0.013	-0.032†	0.019
Constant	-8.431**	1.186	-12.933**	0.841	-11.060**a	1.212
γ_i^{-1} (dispersion parameter)	0.278**	0.029	0.273**	0.072	0.260**	0.027
Log-likelihood at convergence	-2,354.2		-2,507.7		-1,105.8	
AIC	4,730.4		5,037.4		2,233.6	
BIC	4,782.2		5,089.3		2,273.3	
Sample size	274		822		274	

** Significant at 99% level; * Significant at 95% level; † Significant at 90% level.

We also have test the presence of serial correlation in the residuals (error) from our REPG model. The test is used to show that the residuals follow a specific pattern. We selected Durbin Watson test for panel data as one of the popular tests for this purpose (see Bhargava et al., 1982).

$$DW_i = \frac{\sum_{q=1}^{Q}\left\{\sum_{t=2}^{T}\left(\varepsilon_{itq} - \varepsilon_{it-1q}\right)^2\right\}}{\sum_{q=1}^{Q}\left\{\sum_{t=1}^{T}\left(\varepsilon_{itq}\right)^2\right\}}, and \qquad (6\text{-}8)$$

$$\varepsilon_{itq} = k_{itq} - \lambda_{itq}^{REPG}$$

The DW value is distributed in the $[0,4]$ range, with an expected value two indicating no first order serial correlation. When the successive residuals ε_{itq} are close to each other the test value of DW will be low, indicating a presence of a positive serial correlation that we are concerned about. The DW tested values are 0.12, 0.13, 0.30, 0.70 and 0.78 for Total crash, rear end, sideswipe fixed object and all-other types respectively, which are less than the corresponding critical values, indicate to accept the null hypothesis with a statistical evidence that the residuals are positively serially correlated under 1% significance level. This is clearly shown in Figure (6-1) and Figure (6-2), as we plotted the residuals that get from the REPG model of each segment vary with the time of observation for each crash type and the total crash count.

We discuss now the parameters of the crash sum model for the total crash outcome. The parameters are shown in the far right column of Table (6-1) as shown in Table (6-1). The logarithm of ADT and logarithm of length are positively signed, while, the urban-rural indicator, diamond interchange indicator, shortest vertical curve length, and largest rate of vertical curvature are negatively signed. Cross sections with greater than 3 lanes, number of horizontal curves and gradient of shortest vertical curve in segment are positively signed.

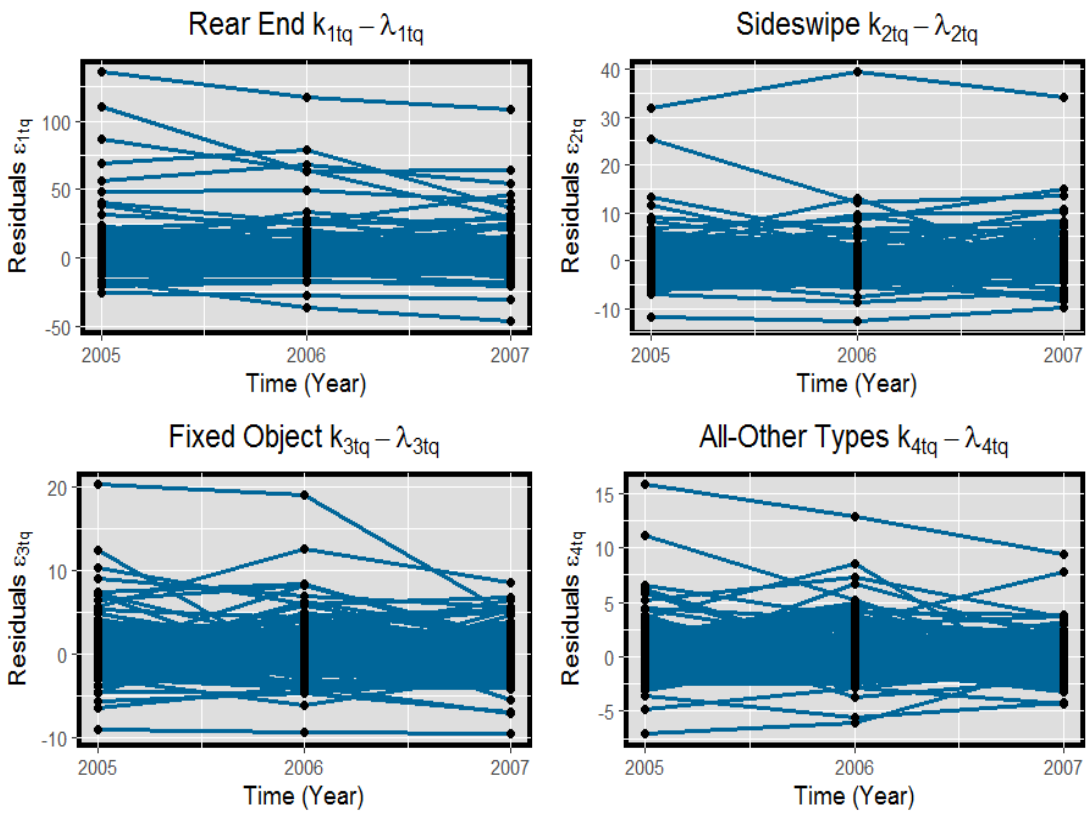


Figure (6-1) Residuals vary with time for each crash type

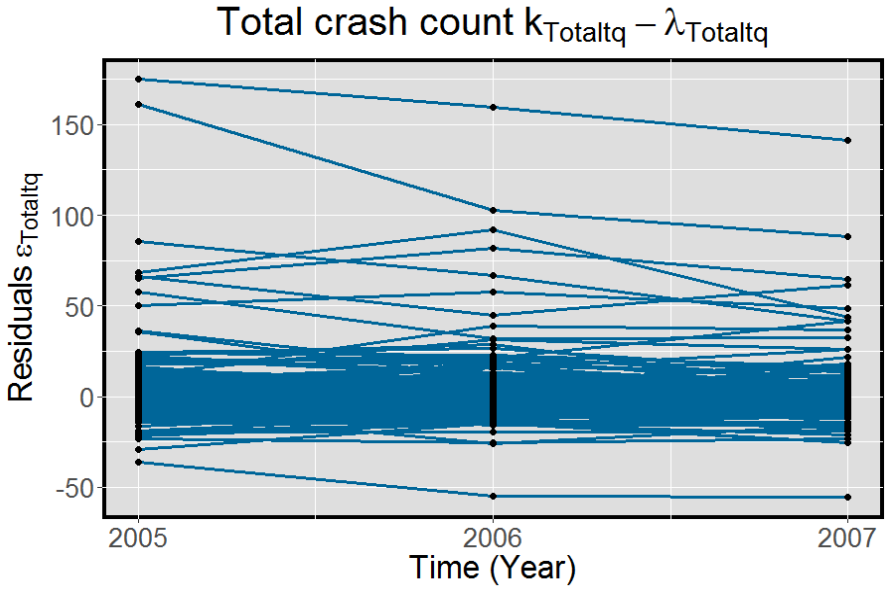


Figure (6-2) Residuals vary with time for the total crash count

The parameter effect sizes are quite similar to that of the REPG model. The constant value in the crash sum model reflects the adjustment due to the time duration effect described in Eq. (6-6). The dispersion parameter is significant in the crash sum model and with a comparable magnitude to the REPG model. The standard error of the dispersion parameter is similar as well. Notably, the standard errors are highly similar, with variations within 7% at the maximum. A majority of the parameters in the crash sum model have standard errors within +/- 2 percent of the REPG model. This pattern of parameter behavior repeated for the individual crash type models as well, which we do not discuss in this paper. To briefly summarize the parameters that were significant in the individual crash type models, such as rear end, sideswipe, and fixed objects included the logarithm of ADT, logarithm of length, the urban-rural indicator, the proportion of segment with three or more lanes variable, the number of horizontal curves variable and the diamond interchange indicator variable. The all-other crash type model included the logarithm of ADT, logarithm of length, and proportion of segment with three or more lanes. The standard errors (as shown in Table (6-2)) of the parameters in the individual crash type models were within +/- 7% of the REPG model of the same crash type. The only exception was the standard error of the dispersion parameter which varied as much as +/- 20% (for rear-end) compared to the REPG model.

6.6 SUMMARY

This study evaluated the random effects Poisson gamma model against a crash sum model for a three-year panel of crash counts in Washington State for the period 2005-2007. We found that the REPG model and the crash sum model were empirically equivalent based on tests of parameter similarities. The crash sum model appears to be a reasonable empirical alternative to the panel model given that the standard errors and parameter magnitudes are highly similar in the presence of time invariant heterogeneity. The time invariant gamma heterogeneity effect captured in the REPG indicates that it is equivalent to an NB distribution for a dependent variable that is equal to the sum of crashes in the whole period, and with an expectation equal to the sum of expectations from the individual time periods. The reasonableness of the crash

sum model as an alternative appeals from the fact that it requires less frequent measurements of geometrics and ADT. This is under the assumption that dramatic changes in the independent variable vector do not occur in the panel observation period. Since we tested a three-year panel, the conclusion of what constitutes a time threshold for the application of the crash sum model is indeterminate. In the crash sum model, the assumption of a representative **X** value will have an impact on the parameter outcomes if there are significant changes to the **X** value over multiple periods. Since, in our dataset, the changes in **X** were minimal, the use of an average value versus an individual year value made little difference. However, this assumption cannot be validated in longer time periods when design changes start to have a noticeable effect. This aspect of constancy of **X** needs to be researched in longer panels before further conclusions on the suitability of the crash sum model can be made. The added benefit of the crash sum model however, is that it approximates the time invariant heterogeneity effect on parameter estimates reasonably well in crash contexts where the crash distribution is from roadways with high exposure (such as the interstate system in our case study). It remains to be seen as to how the model behavior changes with respect to low exposure part of the highway network, and parts of the network where the **X** vector can change frequently (such as variable message signing areas). Nevertheless, the capture of time invariant heterogeneity via the crash sum model ensures that the appropriate variables are identified as statistically significant in the model. This allows for proper model specification across multiple crash type outcomes.

Table (6-2) Three-year crash sum model standard error ratios with respect to REPG model.

Explanatory Variable	Total	Rear-End	Sideswipe	Fixed Object	All-Other
LnAADT	1.023	0.998	1.026	1.013	1.005
LnLength	0.977	0.937	1.001	1.001	1.000
Urban rural dummy, 1 if rural, 0 if urban	0.983	0.967	1.006	1.002	-
Proportion of three or more lanes cross section	0.976	0.950	1.002	1.001	1.000
Number of horizontal curves per segment	0.973	0.941	1.002	1.000	-
Diamond interchange type dummy	0.983	0.966	1.004	1.001	-
Smallest vertical curve gradient in segment	0.978	-	1.001	-	-
Largest beginning vertical curve elevation	-	-	1.002	-	-
Longest horizontal curve central angle	-	-	-	1.000	-
Number of vertical curves	-	-	-	1.000	-
Shortest vertical curve length	0.976	-	-	-	-
Largest vertical curve rate of vertical curvature	0.968	-	-	-	-
Constant	1.022	0.996	1.026	1.013	1.005
γ_i^{-1} (dispersion parameter)	0.925	0.814	1.001	0.998	0.995
Log-likelihood at convergence	-856.6	-669.8	-455.5	-493.9	-449.6
AIC	1,735.1	1,355.6	931.0	1,007.7	909.1
BIC	1,774.9	1,384.5	967.1	1,043.8	927.2
Sample size	274	274	274	274	274

Ratio = standard error of crash sum model/ standard error of REPG model.

Chapter 7

Multivariate Panel Copula-Based Model

7.1 INTRODUCTION

This chapter suggests an econometric framework to model the multivariate panel crash by type count data. The point of emphasis is that modeling multivariate panel count data has more superior econometric benefits, which is clarified in producing more efficient parameter estimates compared to the ones arise from the multivariate cross-sectional models. Within this context, we considered the intertemporal correlations of a given crash type among the years of observations. Moreover, we have considered the inter-type correlations that formulated from jointing the probability among different crash types. Both of these correlation components are added more intricacy to seek a conceivable inference. We developed two flexible models to overcome this problem: Multivariate Panel Poisson Gamma Copula (MVPPGC) and Multivariate Panel Copula-Copula model (MVPCC). These two models are in no need for a simulation mechanism, which is a common issue to model the multivariate count outcomes. The source of flexibility of these models are demonstrated through allowing a non-linear asymmetric shape of these correlation components that generated among the unobserved heterogeneity of each crash type and across the years of observations.

7.2 BACKGROUND

Panel count data or longitudinal count data are observations arise from the count process over a specific time period. Crash panel count data with no exception is a part of nature of this process which is observed for a specific roadway individual entity (segment/intersection) over a one year usually. Crash

panel count data are more elaborate and complex to deal with, because the same individual entity is observed over time. As it is pointed out in Camerion and Trivedi (2013), a key advantage of modeling panel count data over cross-section count data is that the former allows more general types of individual heterogeneity, as we will see later. Most of the developed multivariate models are cross-section data models in the literature. The nature of the complexity of both the multivariate and panel data alongside with the counting process are still to be considered one of the biggest challenges in econometric modeling (Camerion and Trivedi, 2013). The multivariate panel count[1] data emerge in crash count data involving various crash types in which these types are measured over a specific period of time on a specific location. Panel count data itself refers to multidimensional data, other dimensions come from joint each crash count type together. On bivariate/multivariate cross-section count models the reader can refer to Imprialou and Quddus (2016); Lee et al., (2015); Li et al., (2015); Park and Lord (2007). Alternatively, on univariate panel crash count models where the time of observation plays an important role in estimation can be found in (on the fixed effect see Law et al., 2009; Kumara and Chin, 2004; on random effect Poisson or negative binomial (NB) see Chin and Quddus, 2003 and Quddus, 2008). There is a bigger picture unfolding with respect the relationship among different crash count outcomes in panel data framework that needs to be considered. Recent advancements in econometric modeling techniques have allowed researchers to extend univariate crash panel data analysis to a higher dimension and generate a well-known class of multivariate panel crash count models. The latter models outperform the corresponding univariate ones and permit more dynamic dependency through the unobserved heterogeneity. Furthermore, a comparison among different crash count outcomes can be achieved through draw inferences from the joint analysis of the multiple response count variables (Katuwandeniyage and

[1] A comprehensive review on multivariate count models and panel count regression models can be found in Cameron and Trivedi (2013) chapter (8), chapter (9) respectively. in addition, another extensive review can be found in Winkelmann (2013), on correlated count data see Chapter (7) includes a subsection (7.2) on panel data models.

Priyantha, 2015). Unsurprisingly, multivariate panel crash count models are more attractive than simply univariate panel models. Multivariate crash count models still dominate the researchers' main topics in the literature (see Ma et al., 2008). In the context of multivariate panel crash count models, the state of the art in estimating the interdependence of multiple traffic safety outcomes over the years of observations involves simulation based parameter estimation. Generally speaking, any multivariate cross-section crash count model can be transformed into a univariate panel count model (Winkelmann, 2013); for example, the multivariate negative binomial count model becomes random effect Poisson-gamma mixture (also known as the multinomial negative binomial count model). Bhat et al. (2014a) addressed three major types of multivariate cross-section count data approaches as we saw in Chapter 5)). Recall 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, 2013). Our approach is relatively close to the latent Poisson-normal model but with more extension, and in no need to a simulation solution.

A certain belief that the correlation among the unobserved heterogeneity for the same crash type among the years of observation and among the crash types cause serious efficiency problems has been entrenched in literature (see for example Ulfarsson and Shankar, 2003; Sittikariya et al.,2005). As a consequence of formulating the joint probability, both of these correlation components are added more difficulty to seek a conceivable inference. To include the intertemporal correlations into the first category, the fixed/random effect is a common concept to accommodate the individual effects among the years of observations.

Both the bivariate copula function and the CML technique for count data are explained extensively in **Chapter** 5). In the following sections, we will investigate the actuarial-related modeling problems that associated with the multivariate panel crash count models context, and apply our proposed methodology.

7.2.1 Unobserved heterogeneity among years-crash types

A common issue associated with panel count model is the correlation among the years of observation also known in the literature as a serial correlation, autoregressive, intertemporal correlation effect problem. The intertemporal correlation (here and after) occurs when the unobserved heterogeneity (unobserved factors) are correlated over the years of observations (Ulfarsson and Shankar, 2003). The intertemporal correlation among time periods will not affect the unbiasedness or consistency of the model parameters, but it does affect their efficiency (Greene, 2003). A positive intertemporal correlation is noticeable in estimated panel count models with standard errors smaller than the corresponding values produced by the cross-section models. Therefore, it's imperative to model the intertemporal correlation issue to avoid a wrong conclusion on interpreting the parameter estimates that seem more accurate than they really are, which means the chance to include an insignificant variable in the model.

Panel count data give an opportunity to distinguish between the true and apparent contagion[1] issue for the specific individual. This feature allows more control of unobserved heterogeneity on count data modeling for the specific individual (Cameron and Trivedi, 2013). For example, controlling for a specific individual freeway segment (our case context) propensity to produce a certain number of crashes. For a single cross-section, these controls can only rely on the actual observed explanatory variables such as the physical characteristics of the segment, and estimates may become inconsistent if there is additionally an unobserved component to the specific individual segment. Within multivariate panel data, one can additionally include a term which represents

[1] Given the background of count data, If the composition of the observation unit (segment) changes over consecutive trails (crash occurrence is a success trail), as far as, it happens to exist three associated scenarios that appear in count model assumptions. These scenarios are: First, occurrence-dependence, which the composition changes as a consequence of previous success trail. Second, duration-dependence, which occurs as a consequence of previous non-success. Third, and finally, a non-stationarity, which the composition changes due to external reasons independently from the pervious process (fail/success). Both of first and second scenarios are known as a 'contagion' in the statistics literature (Winkelmann, 2013). Contagious situation violates the Poisson distribution equality assumption which is the common case for accident occurrence predisposition

the specific individual segment for the unobserved heterogeneity that is time-varying to accommodate the intertemporal along with/without a different separation parameter to include the time-invariant unobserved heterogeneity such as the ones that causes the over-dispersion problem.

7.3 MODELING FRAMEWORK

In this section, at first, we will formulate the multivariate panel count models based on the random effect approach using the Poisson-gamma mixture to accommodate the intertemporal correlations among the years of observation and the copula function will be used to join the crash types simultaneously. For this type of model, we have selected one parameter to reflect both the temporal correlation and the time-invariant (overdispersion) effect. The second part is devoted to more flexible model that we use copula function to hold the intertemporal correlation instead the random effect and separate the parameter of the overdispersion from the serial correlation. Let *i* $(i=1,2,....,I)$ be an index of the *i*th observed crash type. Let also assume *t* $(t=1,2,....,T)$ an index of unit of time of observation in the panel crash record. Let *q* $(q=1,2,....,Q)$ represents an index of the observation unit (number of segments of the interstate freeway).

7.3.1 Multivariate Mixture Panel Count Model

The mixture distribution function is one of possible ways to construct the multivariate panel count model. Let a count outcome variable y_{itq} can take the value m_{itq}, (where m_{itq} may take one of the positive integer number; i.e., $m_{itq} \in \{0,1,2,....,\}$), which is the observed crash count of crash type *i* at the observed time *t* for the segment *q*. The multiplicative individual (segment) specific effect is used to represent the error components interaction in the expected crash count function as (Cameron and Trivedi, 2013)

$$\begin{aligned} E\left[y_{itq} \mid \mathbf{x}_{itq}, \boldsymbol{\beta}_i, \upsilon_{itq}, \alpha_{itq}\right] &= m_{itq} = \eta_{itq} \\ &= \lambda_{itq} \times \alpha_{itq} \times \upsilon_{itq} \\ &= \exp\left(\boldsymbol{\beta}_i' \mathbf{x}_{itq}\right) \times \alpha_{itq} \times \upsilon_{itq} \end{aligned} \tag{7-1}$$

where, $\mathbf{x}_{itq}$ is a $(H \times 1)$-vector of explanatory variables (including a constant),

$\boldsymbol{\beta}_i$ is an individual specific ($H \times 1$)-vector of parameter to be estimated. α, is a random variable represent the intertemporal correlation among the years of observations. υ, is a random variable represent the correlation among the crash types. Since, $y_{itq} \sim \Pr\left[\lambda_{itq}\alpha_{itq}\upsilon_{itq}\right]$, using the mixture distribution function, we can write the joint probability of the count outcome varies for different crash type on different year of observation as,

$$\Pr\left(y_{itq}\right) = \int_{\forall \upsilon_{itq}} \dots \int_{\forall \alpha_{itq}} \left[\prod_{i=1}^{I}\prod_{t=1}^{T} \Pr\left(y_{itq} \mid \lambda_{itq}, \alpha_{itq}, \upsilon_{itq}\right)\right] f_{Mq}\left(\boldsymbol{\alpha}_q, \boldsymbol{\upsilon}_q\right) d\boldsymbol{\alpha}_q d\boldsymbol{\upsilon}_q \qquad (7\text{-}2)$$

where, f_{Mq} is a multivariate density function that hold the unobserved heterogeneities components. Let also two indices ($h = 1,2,\dots,T$) and ($j = 1,2,\dots,I$), where the dimensions of $\boldsymbol{\alpha}_q$ is not mandatory to equal to $\boldsymbol{\upsilon}_q$, and both are given as[1],

$$\alpha_{itq}, \upsilon_{itq} \sim f_{Mq}\left[\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}; \begin{pmatrix}\Gamma_{i11} & & & & \\ \varsigma_{i21} & \Gamma_{i22} & & & \\ \varsigma_{i31} & \varsigma_{i32} & \Gamma_{i33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \varsigma_{iT1} & \varsigma_{iT2} & \cdots & \varsigma_{iTT-1} & \Gamma_{iTT}\end{pmatrix}_{\forall i}, \begin{pmatrix}\Gamma_{11t} & & & & \\ \Omega_{21t} & \Gamma_{22t} & & & \\ \Omega_{31t} & \Omega_{32t} & \Gamma_{33t} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \Omega_{I1t} & \Omega_{I2t} & \cdots & \Omega_{II-1t} & \Gamma_{IIt}\end{pmatrix}_{\forall t}\right]_q$$

$$\text{Or} \qquad (7\text{-}3)$$

$$\alpha_{itq}, \upsilon_{itq} \sim f_{Mq}\left[(\mathbf{0},\mathbf{0}); \mathbf{VT}, \mathbf{VI}\right]$$

where $\mathbf{VT}, \mathbf{VI}$ are the varaince-covaraince matrices of the unobserved heterogeneities generated from each correlation of the error terms for each crash years-types with equal variances property when $\mathbf{VT}_{\mathrm{i\,(t,t)}} = \mathbf{VI}_{\mathrm{(i,i)\,t}}$ and unequal covariances property when $\mathbf{VT}_{\mathrm{i\,(h,t)}} \neq \mathbf{VI}_{\mathrm{(j,i)\,t}}$. The conjugated

[1] Normalizing the error components fundamentally depend on the selection of the mixture function. For example, for a univariate single share random effect count model for the serial correlation, if the gamma distribution (mixture function) is selected to represents the unobserved heterogeneity, the error will be normalized to 1 instead of 0 (that's also applicable for the integral domains, which in that case it would be (0~+inf)).

probabilities represented by Eq. (7-2) of the unconditional crash count probability has no closed-form solution. The solution using numerical integration is a computationally intensive (see Cameron and Trivedi, 2013; Paleti and Bhat, 2013). We overcome this difficulty by introducing a simple formula that can approximate the probability of Eq. (7-2) using altogether the random effect concept, copula function and the CML method. The multivariate panel count model by definition assumes the explanatory effects are identical and consistent across the multiple time periods.

7.3.1.1 Multivariate panel Poisson-gamma mixture-copula count model (MVPPGC):

As we mentioned earlier, in this type of model, we will let the Poisson-gamma mixture model to hold the probability of the individual crash type across the years of observations to account the intertemporal correlation problem, while the correlation among the crash types itself will be represented through the copula function and the CML technique. Modeling the intertemporal correlation using Poisson-gamma mixture model is well known in the literature (as a random effect model or multinomial Negative binomial, random effect Poisson gamma mixture …etc). It is usually started by letting the unobserved heterogeneity random variable α (as a single share property) to vary across the segments for each crash type and being subjected to a certain univariate continues distribution $f(\alpha_{iq})$ (we will use gamma distribution, but any i.i.d continuous distribution is also suitable) (Camerion and Trivedi, 2013). Eq. (7-2) collapses into,

$$\Pr\left(y_{i1q}, y_{i1q}, ..., y_{iTq}\right) = \prod_{i=1}^{I}\left\{\int_{0}^{+\infty}\left[\prod_{t=1}^{T}\Pr\left(y_{itq} \mid \lambda_{itq}, \alpha_{iq}\right)\right] f\left(\alpha_{iq}\right) d\alpha_{iq}\right\} \tag{7-4}$$

Let's assume the observed number of crashes m_{itq} is drawn from a Poisson distribution as,

$$\begin{aligned} &\left(y_{itq} = m_{itq}\right) \sim P\left[\eta_{itq} = \alpha_{iq} \times \exp\left(\boldsymbol{\beta}_i' \mathbf{x}_{itq}\right)\right] \\ &P\left[y_{itq} = m_{itq}\right] = \frac{e^{-\eta_{itq}} \times \eta_{itq}^{mitq}}{m_{itq}!} \end{aligned} \tag{7-5}$$

Substitute Eq. (7-5) into Eq. (7-4) and the likelihood function has a closed-form solution as,

$$L[\boldsymbol{\beta}_i,\boldsymbol{\gamma}]=\prod_{i=1}^{I}\left\{\frac{\gamma_i^{\gamma_i}}{\Gamma(\gamma_i)}\left(\prod_{t=1}^{T}\frac{\lambda_{itq}^{m_{itq}}}{m_{itq}!}\right)\int_0^{\infty}\exp\left(-\alpha_{iq}\sum_{t=1}^{T}\lambda_{itq}\right)\alpha_{iq}^{\sum_{t=1}^{T}m_{itq}}\alpha_{iq}^{\gamma_i-1}\exp\left(-\gamma_i\alpha_{iq}\right)d\alpha_{iq}\right\}$$

$$=\prod_{i=1}^{I}\left\{\frac{\gamma_i^{\gamma_i}}{\Gamma(\gamma_i)}\left(\prod_{t=1}^{T}\frac{\lambda_{itq}^{m_{itq}}}{m_{itq}!}\right)\int_0^{\infty}\exp\left(-\alpha_{iq}\left[\gamma_i+\sum_{t=1}^{T}\lambda_{itq}\right]\right)\alpha_{iq}^{\gamma_i+\sum_{t=1}^{T}m_{itq}-1}d\alpha_{iq}\right\} \quad (7\text{-}6)$$

$$=\prod_{i=1}^{I}\left\{\left(\prod_{t=1}^{T}\frac{\lambda_{itq}^{m_{itq}}}{m_{itq}!}\right)\frac{\Gamma\left(\gamma_i+\sum_{t=1}^{T}m_{itq}\right)}{\Gamma(\gamma_i)}\left(\frac{\gamma_i}{\gamma_i+\sum_{t=1}^{T}\lambda_{itq}}\right)^{\gamma_i}\left(\frac{1}{\gamma_i+\sum_{t=1}^{T}\lambda_{itq}}\right)^{\sum_{t=1}^{T}m_{itq}}\right\}$$

where $\boldsymbol{\beta}_i=(\beta_{i1},\beta_{i2},.....,\beta_{iH})^T$, $\boldsymbol{\gamma}=(\gamma_1,\gamma_2,.....\gamma_I)$ and $\alpha_{iq}\overset{i.i.d.}{\sim}G(\gamma_i,\gamma_i)$ with expected mean $E(\alpha_{iq})=1$ and variance $1/\gamma_i$. Eq. (7-6) is the multiplayers of the random effect Poisson gamma mixture probabilities across all the crash types. This distribution has suggested by Hausman et al (1984) with the $E(y_{itq})=\lambda_{itq}$ and $V(y_{itq})=\lambda_{itq}+\lambda_{itq}^2/\gamma_i$ properties. The time invariant unobserved heterogeneity and the intertemporal correlation are represented by same dispersion parameter γ_i.

We will select the bivariate copula function with the following properties $C[F_1(y_{1tq}),0]=C[0,F_2(y_{2q})]=0$; $C[F_1(y_{1q}),1]=F_1(y_{1q})$ and $C[1,F_2(y_{2q})]=F_2(y_{2q})$, which allows us to solve the integral of the joint distribution and to seek for a non-linear and asymmetric patterns of relationships among the error terms which give more flexibility in modeling for more details see (Mothafer et al., 2016; Bhat and Eluru, 2009). The model estimation is carried out after specifying a suitable marginal distribution *F* for the count outcome and an appropriate copula *C*. The CML technique has been utilized to overcome the multi-dimensionality that generated from the dependencies among the crash types without a need to evaluate the full likelihood function given in Eq. (7-2) (Bhat et al., 2014b; Bhat et al.,2014c; Castro et al., 2012; Castro et al., 2013; Yamamoto and Morikawa;

2013; Sener et al., 2010; Ferdous et al., 2010; Paleti and Bhat, 2013). Let the observations for a given year as $\left(m_{1tq}, m_{2tq}, ..., m_{Itq}\right)$. The pairwise CML which works with the bivariate copula perfectly, is used to obtain the joint probability among the crash types, thus the latter expression collapses into $T \times (I \times (I-1)/2)$ pairs of bivariate probability computations and it takes the form,

$$
\begin{aligned}
L_{CML_q}^{Crash}\left[\boldsymbol{\beta}_i, \boldsymbol{\gamma}, \boldsymbol{\theta}\right] &= \prod_{t=1}^{T}\left\{\prod_{i=1}^{I-1}\prod_{j=i+1}^{I}\Pr\left(y_{tiq}=m_{itq}, y_{jtq}=m_{jtq}\right)\right\} \\
&= \prod_{t=1}^{T}\left\{\prod_{i=1}^{I-1}\prod_{j=i+1}^{I}\left[\int_{F_{it}(m_{itq}-1)}^{F_{it}(m_{itq})}\int_{F_{jt}(m_{jtq}-1)}^{F_{jt}(m_{jtq})} c_{(i,j)t}\left[F_{it}(y_{itq}), F_{jt}(y_{jtq}) \mid \theta_{ij}\right].\prod_{i=1}^{I} f_{it}(y_{itq})\, dF_{it}(y_{itq}) dF_{jt}(y_{jtq})\right]\right\} \quad (7\text{-}7) \\
&= \prod_{t=1}^{T}\left\{\prod_{i=1}^{I-1}\prod_{j=j+1}^{I}\left[\begin{array}{l} C\left[F_{it}(m_{itq}), F_{jt}(m_{jtq}) \mid \theta_{ij}\right] - C\left[F_{it}(m_{itq}-1), F_{jt}(m_{jtq}) \mid \theta_{ij}\right] \\ -C\left[F_{it}(m_{itq}), F_{jt}(m_{jtq}-1) \mid \theta_{ij}\right] + C\left[F_{it}(m_{itq}-1), F_{jt}(m_{jtq}-1) \mid \theta_{ij}\right] \end{array}\right]\right\}
\end{aligned}
$$

And we specify $\boldsymbol{\theta} = \begin{bmatrix} \theta_{21} & & & \\ \theta_{31} & \theta_{32} & & \\ \vdots & \vdots & \ddots & \\ \theta_{I1} & \theta_{I2} & \cdots & \theta_{II-1} \end{bmatrix}$ (7-8)

where θ_{ij} represents the level of dependency between the marginals $F_{it}(m_{itq}), F_{jt}(m_{jtq})$ for a certain copula function *C* (there is a $(I \times (I-1)/2)$ θ_{ij} parameters in total). The marginal distribution that we selected is the cumulative negative binomial count distribution NBII, which accommodates the time invariant unobserved heterogeneity for each given year of observation of each individual crash type as,

$$
\begin{aligned}
\Pr\left(y_{itq} \le m_{tiq} \mid \mathbf{x}_{itq}, \boldsymbol{\beta}_i\right) = F_{it}\left(m_{itq}\right) &= \sum_{r=0}^{m_{itq}} f_{it}\left(r \mid \mathbf{x}_{itq}, \boldsymbol{\beta}_i\right) \\
&= \sum_{r=0}^{m_{itq}}\left[\frac{\Gamma\left(r+\gamma_i\right)}{\Gamma\left(\gamma_i\right)\Gamma\left(r+1\right)}\left(\frac{\gamma_i}{\lambda_{itq}+\gamma_i}\right)^{\gamma_i}\left(\frac{\lambda_{itq}}{\lambda_{itq}+\gamma_i}\right)^{r}\right]
\end{aligned} \quad (7\text{-}9)
$$

where $\lambda_{itq} = \exp\left(\boldsymbol{\beta}_i' \mathbf{x}_{itq}\right)$, $\Gamma_{it}\left(y_{itq}\right) = \lambda_{itq} + \gamma_i^{-1}\left(\lambda_{itq}\right)^2$ are the conditional mean and the conditional variance respectively (overdispersion occurs when $1/\gamma_i$ > 0).

7.3.1.2 Multivariate panel copula-copula (MVPCC) count model:

In this model, a more flexible structure of the intertemporal correlation among the years of observations is introduced. Rather than assign a parametric distribution for the random effects, a possible alternative is to use the bivariate copula and the pairwise CML, but this time, across the years of observations, then the joint probability in that case are collapses into $I\times(T\times(T-1)/2)$ pairs of bivariate probability computations. Without loss generality, let an index ($h=1,2,...,T$), with the observations among years for a given crash type as $(m_{i1q},m_{i2q},....,m_{iTq})$ so we can write,

$$
\begin{aligned}
L_{CML_q}^{Time}(\boldsymbol{\beta}_i,\boldsymbol{\gamma},\vartheta) &= \prod_{i=1}^{I}\left\{\prod_{t=1}^{T-1}\prod_{h=t+1}^{T}\Pr\left(y_{itq}=m_{itq},y_{ihq}=m_{jhq}\right)\right\} \\
&= \prod_{i=1}^{I}\left\{\prod_{t=1}^{T-1}\prod_{h=t+1}^{T}\left[\int_{F_{it}(m_{itq}-1)}^{F_{it}(m_{itq})}\int_{F_{jt}(m_{jtq}-1)}^{F_{jt}(m_{jtq})}c_{i(t,j)}\left[F_{it}(y_{itq}),F_{jt}(y_{ihq})\mid\vartheta_i\right].\prod_{t=1}^{T}f_{it}(y_{itq})\,dF_{it}(y_{itq})dF_{jt}(y_{ihq})\right]\right\} \quad (7\text{-}10)\\
&= \prod_{i=1}^{I}\left\{\prod_{t=1}^{T-1}\prod_{h=t+1}^{T}\left[\begin{array}{l}C\left[F_{it}(m_{itq}),F_{jt}(m_{jhq})\mid\vartheta_i\right]-C\left[F_{it}(m_{itq}-1),F_{jt}(m_{jhq})\mid\vartheta_i\right]\\-C\left[F_{it}(m_{itq}),F_{jt}(m_{jhq}-1)\mid\vartheta_i\right]+C\left[F_{it}(m_{itq}-1),F_{jt}(m_{jhq}-1)\mid\vartheta_i\right]\end{array}\right]\right\}
\end{aligned}
$$

where ϑ_i is a specified parameter of the selected parametric copula that represents the intertemporal correlation[1] ,which is separated from the time-invariant unobserved heterogeneity that represented by the dispersion parameter γ_i. The count marginal distribution of Eq. (7-10) is also the NBII cumulative distribution as,

$$
\begin{aligned}
\Pr\left(y_{itq}\le m_{tiq}\mid\mathbf{x}_{itq},\boldsymbol{\beta}_i\right) &= F_{it}\left(m_{itq}\right)=\sum_{r=0}^{m_{itq}}f_{it}\left(r\mid\mathbf{x}_{itq},\boldsymbol{\beta}_i\right)\\
&= \sum_{r=0}^{m_{itq}}\left[\frac{\Gamma\left(r+\gamma_i\right)}{\Gamma\left(\gamma_i\right)\Gamma\left(r+1\right)}\left(\frac{\gamma_i}{\lambda_{itq}+\gamma_i}\right)^{\gamma_i}\left(\frac{\lambda_{itq}}{\lambda_{itq}+\gamma_i}\right)^{r}\right] \quad (7\text{-}11)
\end{aligned}
$$

the joint probability among the crash type for the MVPCC model in that case is

[1] There are several structures to represent the variance-covariance of the temporal correlation for a given copula (e.g. independent; autoregressive of order 1 (AR (1)); Toeplitz; banded Toeplitz; unstructured and compound symmetry (also known as exchangeable)). We selected the compound symmetry were all the correlation parameters are constant over time, but another configuration could be also possibly augmented.

also given by the Eq. (7-7) and Eq. (7-8).

Assume three vectors of the parameter estimators of the two models are given as $\zeta_{1q}=(\beta_{i1},\beta_{i2},....,\beta_{iH};\gamma_1,\gamma_2,....,\gamma_I)$,

$\zeta_{2q}=(\beta_{i1},\beta_{i2},....,\beta_{iH};\gamma_1,\gamma_2,....,\gamma_I;\vartheta_1,\vartheta_2,....,\vartheta_I)$

and $\zeta_{3q}=(\beta_{i1},\beta_{i2},....,\beta_{iH};\gamma_1,\gamma_2,....,\gamma_I;\boldsymbol{\theta})$. The first estimator for the MVPPGC model represents the parameters for the years of observations model given by the Poisson gamma mixture for a given crash type, which corresponds to the second estimator for the MVPCC model, except we have additional four intertemporal correlation parameters instead of one. The final estimator is the one represents the parameters in the joint probability among the crash types for a given year of observations which is used for both models.

The likelihood functions for both models are easy to maximize, where the estimators $\zeta_{1,2,3_q}$ are obtained by maximizing the logarithm of the sum of the two parts of likelihood functions for both models. These estimators are consistent and asymptotically normally distributed with asymptotic mean $\zeta_{1,2,3_q}$ and covariance matrix given by the inverse of Godambe's (1960) sandwich information matrix $\mathbf{G}(\zeta)$ (see Zhao and Joe, 2005; Castro et al., 2012 and Ferdous et al. 2010). Define both $\mathbf{V}(\zeta_{1,3})$, $\mathbf{V}(\zeta_{2,3})$ for MVPPGC and MVPCC model respectively. Similarly, define $\mathbf{G}(\zeta_{1,3})$, $\mathbf{H}(\zeta_{1,3})$, $\mathbf{J}(\zeta_{1,3})$ and $\mathbf{G}(\zeta_{1,3})$, $\mathbf{H}(\zeta_{1,3})$, $\mathbf{J}(\zeta_{1,3})$. Where $\mathbf{H}(\zeta)$ and $\mathbf{J}(\zeta)$ are the Hessian and the Jacobian matrices. Starting with the MVPPGC model the score function can be written as,

$$\begin{aligned}\mathbf{V}(\zeta_{1,3}) &= [\mathbf{G}(\zeta_{1,3})]^{-1}\\ &= [\mathbf{H}(\zeta_{1,3})]^{-1}\mathbf{J}(\zeta_{1,3})[\mathbf{H}(\zeta_{1,3})]^{-1}, \text{where}\\ \mathbf{H}(\zeta_{1,3}) &= E\left[-\left(\frac{\partial^2 \ln L_q(\zeta_1)}{\partial\zeta_1\,\partial\zeta_1'}+\frac{\partial^2 \ln L_{CMLq}^{Crash}(\zeta_3)}{\partial\zeta_3\,\partial\zeta_3'}\right)\right] \text{and}\\ \mathbf{J}(\zeta_{1,3}) &= E\left[\left(\frac{\partial \ln L_q(\zeta_1)}{\partial\zeta_1}\right)\left(\frac{\partial \ln L_q(\zeta_1)}{\partial\zeta_1'}\right)+\left(\frac{\partial \ln L_{CMLq}^{Crash}(\zeta_3)}{\partial\zeta_3}\right)\left(\frac{\partial \ln L_{CMLq}^{Crash}(\zeta_3)}{\partial\zeta_3'}\right)\right]\end{aligned} \quad (7\text{-}12)$$

while the $\mathbf{V}(\zeta_{2,3})$ and $\mathbf{G}(\zeta_{2,3})$ for the MVPCC model are given by replacing the

above by,

$$\mathbf{H}\left(\zeta_{2,3}\right)=E\left[-\left(\frac{\partial^2 \ln L_{CMLq}^{Time}\left(\zeta_2\right)}{\partial\zeta_2\,\partial\zeta_2'}+\frac{\partial^2 \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3\,\partial\zeta_3'}\right)\right] \text{ and}$$

$$\mathbf{J}\left(\zeta_{2,3}\right)=E\left[\left(\frac{\partial \ln L_{CMLq}^{Time}\left(\zeta_2\right)}{\partial\zeta_2}\right)\left(\frac{\partial \ln L_{CMLq}^{Time}\left(\zeta_2\right)}{\partial\zeta_2'}\right)+\left(\frac{\partial \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3}\right)\left(\frac{\partial \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3'}\right)\right]$$

So far, the estimation was formulated for only a given segment q, to get the joint probability estimation across all segments, we write for the MVPPGC model,

$$\hat{\mathbf{H}}\left(\hat{\xi}\right)=\frac{1}{Q}\sum_{q=1}^{Q}\left[-\left(\frac{\partial^2 \ln L_q\left(\zeta_1\right)}{\partial\zeta_1\,\partial\zeta_1'}+\frac{\partial^2 \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3\,\partial\zeta_3'}\right)\right]_{\hat{\xi}}$$

$$=\left[\frac{1}{Q}\sum_{q=1}^{Q}\left(\sum_{i=1}^{I}\frac{\partial^2 \ln \Pr\left(y_{itq}=m_{itq}\right)}{\partial\zeta_1\,\partial\zeta_1'}+\sum_{t=1}^{T}\sum_{i=1}^{I-1}\sum_{j=j+1}^{I}\frac{\partial^2 \ln \Pr\left(y_{itq}=m_{itq},y_{jtq}=m_{jtq}\right)}{\partial\zeta\,\partial\zeta'}\right)\right]_{\hat{\xi}},$$

for MVPCC Model

$$=\frac{1}{Q}\sum_{q=1}^{Q}\left[-\left(\sum_{i=1}^{I}\sum_{t=1}^{T-1}\sum_{h=t+1}^{T}\frac{\partial^2 \ln \Pr\left(y_{itq}=m_{itq},y_{jtq}=m_{jtq}\right)}{\partial\zeta_1\,\partial\zeta_1'}\right.\right.$$

$$\left.\left.+\sum_{t=1}^{T}\sum_{i=1}^{I-1}\sum_{j=j+1}^{I}\frac{\partial^2 \ln \Pr\left(y_{itq}=m_{itq},y_{jtq}=m_{jtq}\right)}{\partial\zeta\,\partial\zeta'}\right)\right]$$

Similarly, for the Jacobin matrix for the MVPPGC model

$$\hat{\mathbf{J}}\left(\hat{\xi}\right)=\frac{1}{Q}\sum_{q=1}^{Q}\left[\left[\left(\frac{\partial \ln L_q\left(\zeta_1\right)}{\partial\zeta_1}\right)\left(\frac{\partial \ln L_q\left(\zeta_1\right)}{\partial\zeta_1'}\right)+\left(\frac{\partial \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3}\right)\left(\frac{\partial \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3'}\right)\right]\right]_{\hat{\xi}}$$

and for the MVPCC model

$$\frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\frac{\partial \ln L_{CMLq}^{Time}\left(\zeta_2\right)}{\partial\zeta_2}\right)\left(\frac{\partial \ln L_{CMLq}^{Time}\left(\zeta_2\right)}{\partial\zeta_2'}\right)+\left(\frac{\partial \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3}\right)\left(\frac{\partial \ln L_{CMLq}^{Crash}\left(\zeta_3\right)}{\partial\zeta_3'}\right)\right]_{\hat{\xi}}$$

7.3.2 Model Estimation selection

To assess the performance of our developed models we selected both the Akaike information criterion (*AIC*) and Bayesian information criterion (*BIC*) measures which is applicable for non-nested models (see Nikoloulopoulos and Karlis (2009); Winkelmann (2013); Cameron and Trivedi (2013)). The *BIC* performed better in large samples, whereas the *AIC* tends to be superior in

small samples (Shumway and Stoffer, 2010). The lowest values of these measures indicate a better performance usually. The *AIC* and the *BIC* can be given as $AIC = -2\times\log(LL)+2\times(Q)$, and $BIC = -2\times\log(LL)+(\kappa)\times\log(Q)$, where κ is the number of parameters of the copula model. The comparison between the MVPPGC and the MVPCC models is an interesting point to be investigated but first we need to adjust the likelihood function of the pairwise composite marginal likelihood estimate first. The weight $1/(T_q-1)$ is used for a given segment in the Eq. (7-10) so that the MVPPGC and the MVPCC likelihood estimations are comparable when we use the *AIC* and *BIC* measures (See Bhat, 2011) and we can write.

$$L^{Time}_{CML_q}(\beta_i,\gamma,\vartheta)=\prod_{i=1}^{I}\left\{\frac{1}{(T_q-1)}\prod_{t=1}^{T-1}\prod_{h=t+1}^{T}\begin{bmatrix}C\left[F_{it}(m_{itq}),F_{jt}(m_{jhq})\,|\,\vartheta_i\right]-C\left[F_{it}(m_{itq}-1),F_{jt}(m_{jhq})\,|\,\vartheta_i\right]\\-C\left[F_{it}(m_{itq}),F_{jt}(m_{jhq}-1)\,|\,\vartheta_i\right]+C\left[F_{it}(m_{itq}-1),F_{jt}(m_{jhq}-1)\,|\,\vartheta_i\right]\end{bmatrix}\right\} \quad (7\text{-}13)$$

A non-nested likelihood ratio test is also used (see Ben-Akiva and Lerman, 1985) to give an insight of which model is more statistically significant. The estimation score functions of the log-likelihoods were coded in GAUSS Aptech (2014) and the default BFGS algorithm provided by the maxlik procedure in GAUSS was used for maximizing the score functions.

7.3.3 Variance Covariance Structure of Developed Models

The varaince-covaraince matrices $\mathbf{VT}$, $\mathbf{VI}$ of both the MVPPGC and MVPCC models are formulated from the unobserved heterogeneity for a given segment q where both matrices are squared with dimensions $T\times T$, $I\times I$ respectively. The variances appear along the diagonal and the covariances components in the off-diagonal elements as given below,

$$\mathbf{VT}_{T\times T}=\begin{pmatrix}\Gamma_{i11}&&&&\\ \varsigma_{i21}&\Gamma_{i22}&&&\\ \varsigma_{i31}&\varsigma_{i32}&\Gamma_{i33}&&\\ \vdots&\vdots&\vdots&\ddots&\\ \varsigma_{iT1}&\varsigma_{iT2}&\cdots&\varsigma_{iTT-1}&\Gamma_{iTT}\end{pmatrix}_{\forall i},\ \mathbf{VI}_{I\times I}=\begin{pmatrix}\Gamma_{11t}&&&&\\ \Omega_{21t}&\Gamma_{22t}&&&\\ \Omega_{31t}&\Omega_{32t}&\Gamma_{33t}&&\\ \vdots&\vdots&\vdots&\ddots&\\ \Omega_{I1t}&\Omega_{I2t}&\cdots&\Omega_{II-1t}&\Gamma_{IIt}\end{pmatrix}_{\forall t} \quad (7\text{-}14)$$

The expected covariance between two independent random discrete

variables for the MVPPGM model for the years of observations matrix $\mathbf{VT}$ is given using,

$$\varsigma_{i(t,h)}\left(y_{itq}, y_{ihq}\right)=\frac{\lambda_{itq}\times\lambda_{ihq}}{\gamma_i}. \tag{7-15}$$

while for the MVPCC model, we use the Hoeffding's formula (more details see D'Angelo et al., 2013; Hoeffding, 1940)

$$\begin{aligned}\varsigma_{i(t,h)}\left(y_{itq}, y_{jhq}\right)&=\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\Pr\left(y_{itq}\le r, y_{ihq}\le s\right)-\left[\sum_{r=0}^{\infty}\Pr\left(y_{itq}\le r\right)\right]\times\left[\sum_{s=0}^{\infty}\Pr\left(y_{ihq}\le s\right)\right]\\&=\left\{\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}C\left[F_{it}(r), F_{ih}(s)\,|\,\vartheta_i\right]-\left[\sum_{r=0}^{\infty}F_{it}(r)\right]\times\left[\sum_{s=0}^{\infty}F_{ih}(s)\right]\right\}.\end{aligned} \tag{7-16}$$

for both models the covariances components in the $\mathbf{VI}$ matrix among the crash types is given as,

$$\begin{aligned}\Omega_{(i,j)t}\left(y_{itq}, y_{jtq}\right)&=\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}\Pr\left(y_{itq}\le r, y_{jtq}\le s\right)-\left[\sum_{r=0}^{\infty}\Pr\left(y_{itq}\le r\right)\right]\times\left[\sum_{s=0}^{\infty}\Pr\left(y_{jtq}\le s\right)\right]\\&=\left\{\sum_{r=0}^{\infty}\sum_{s=0}^{\infty}C\left[F_{it}(r), F_{jt}(s)\,|\,\theta_{ij}\right]-\left[\sum_{r=0}^{\infty}F_{it}(r)\right]\times\left[\sum_{s=0}^{\infty}F_{jt}(s)\right]\right\}.\end{aligned} \tag{7-17}$$

where $\varsigma_{\mathrm{i\,(h,t)}}\neq\Omega_{\mathrm{(j,i)\,t}}$ and the average of the expected covariance $\varsigma_{i(t,h)}$ and $\Omega_{i(t,h)}$ among all segments are calculated using,

$$\begin{aligned}E\left[\varsigma_{i(t,h)}\left(y_{itq}, y_{ihq}\right)\right]&=\frac{1}{Q}\sum_{q=1}^{Q}\varsigma_{i(t,h)}\left(y_{itq}, y_{ihq}\right)\\E\left[\Omega_{(i,j)t}\left(y_{itq}, y_{jtq}\right)\right]&=\frac{1}{Q}\sum_{q=1}^{Q}\Omega_{(i,j)t}\left(y_{itq}, y_{jtq}\right)\end{aligned} \tag{7-18}$$

and, the total covariances of crash types are also given as,

$$\begin{aligned}Cov^{Time}\left(y_{it}, y_{jh}\right)&=Cov\left(\lambda_{it}, \lambda_{ih}\right)+E\left[\varsigma_{i(t,h)}\left(y_{it}, y_{ih}\right)\right]\\Cov^{Crash}\left(y_{it}, y_{jt}\right)&=Cov\left(\lambda_{it}, \lambda_{jt}\right)+E\left[\Omega_{(i,j)t}\left(y_{it}, y_{jt}\right)\right]\end{aligned} \tag{7-19}$$

The variance components along the diagonals of both the $\mathbf{VT}$, $\mathbf{VI}$ of both the MVPPGC and the MVPCC models are equal when $\mathbf{VT}_{\mathrm{i\,(t,t)}}=\mathbf{VI}_{\mathrm{(i,i)\,t}}$ and both given by, $\Gamma_{it}\left(y_{itq}\right)=\lambda_{itq}+\gamma_i^{-1}\left(\lambda_{itq}\right)^2$ (for NBII marginal distribution). The average of the variances $\Gamma_{itq}\left(y_{itq}\right)$ of all segments is calculated using,

$$E\left[\Gamma_{it}\left(y_{it}\right)\right]=\frac{1}{Q}\sum_{q=1}^{Q}\Gamma_{tiq}\left(y_{itq}\right) \tag{7-20}$$

The total variance magnitude $V_i^{Time}\left(y_{it}\right)$ for the given crash type among all years of observations is obtained from sum of two components respectively as,

$$V_i^{Time}\left(y_{it}\right)=V_i\left[E\left(y_{it}^{Time}\right)\right]+E\left[V_i\left(y_{it}^{Time}\right)\right] \tag{7-21}$$

where the $V_i\left[E\left(y_{it}^{Time}\right)\right]$ represents the variance of the expected number of total crash which is constructed from the observed heterogeneity while the second component $E\left[V_i\left(y_{it}^{Time}\right)\right]$ is the expected variance formulated from the unobserved heterogeneity given in the $\mathbf{VT}$ matrix, both components are given by Eq. (24) respectively.

$$\begin{aligned} V_i\left[E\left(y_{it}^{Time}\right)\right]&=\left[\sum_{t=1}^{T}Var\left(\lambda_{it}\right)+2\sum_{t=1}^{T-1}\sum_{h=t+1}^{T}Cov\left(\lambda_{it},\lambda_{ih}\right)\right]_{\forall i}\\ E\left[V\left(y_{it}^{Time}\right)\right]&=\left[\sum_{t=1}^{T}E\left[\Gamma_{it}\left(y_{it}\right)\right]+2\sum_{t=1}^{T-1}\sum_{h=t+1}^{T}E\left[\varsigma_{i(t,h)}\left(y_{it},y_{ih}\right)\right]\right]_{\forall i}\end{aligned} \tag{7-22}$$

The same goes to the total variance $V_t^{Crash}\left(y_{it}\right)$ for the given year of observation across all crash types

$$V_t^{Crash}\left(y_{it}\right)=V_t\left[E\left(y_{it}^{Crash}\right)\right]+E\left[V_t\left(y_{it}^{Crash}\right)\right] \tag{7-23}$$

analogous to previous

$$\begin{aligned} V_t\left[E\left(y_{it}^{Crash}\right)\right]&=\left[\sum_{i=1}^{I}Var\left(\lambda_{it}\right)+2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}Cov\left(\lambda_{it},\lambda_{jt}\right)\right]_{\forall t}\\ E\left[V\left(y_{it}^{Crash}\right)\right]&=\left[\sum_{i=1}^{I}E\left[\Gamma_{it}\left(y_{it}\right)\right]+2\sum_{i=1}^{I-1}\sum_{j=i+1}^{I}E\left[\Omega_{(i,j)t}\left(y_{it},y_{jt}\right)\right]\right]_{\forall t}\end{aligned} \tag{7-24}$$

7.4 EMPIRICAL CRASH DATA SETTING

7.4.1 Configuration

Here, also the crash-count record is considered as panel data. The crashes sample size produces (274) segment-year observations for (3) years of observation for each crash type. The crash-count type distributions are

presented in Figure (3-2) while the descriptive statistic of the main explanatory variables in this study is shown in Table (3-2).

7.4.2 Temporal correlations patterns in the crash data

A multiple time series plots of each crash type in the crash count data over three years of observations (from 2005 to 2007) is presented in Figure (7-1) In these plots the heterogeneity among segments, observations within a specific segment over the observation period tend to have the same value as compared to observations across segments for a given year. These plots also indicate a strong segment effects through a highly temporal correlation for crash type in our crash count data (temporal patterns are evident).

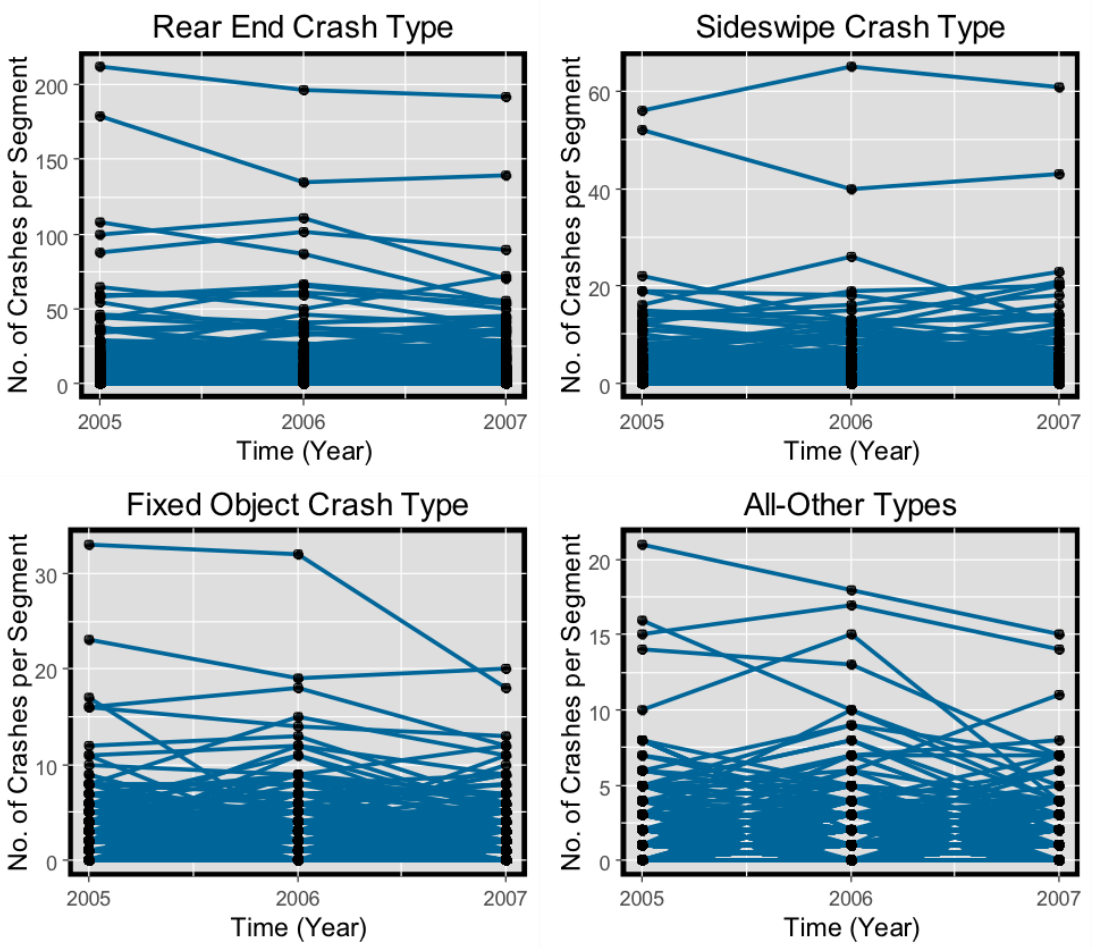


Figure (7-1) Multiple time series plot of different crash type counts from 2005 to 2007.

7.5 MODEL ESTIMATION AND PERFORMANCE

In this section, we will mainly implement the developed models into our crash type's dataset. But before that, we took the chance to take some results from our existed work (see Mothafer, et al. 2016) to select which type of copula we can start with to get maximum performance in a sense of offering better goodness-of-fit. Based on this work, for our both the MVPPGC and the MVPCC developed models we used and fixed "Frank" copula function among crash types for a given year of observation, since the former offers better fits among all the other copula functions. As for the MVPCC model we need to investigate also which copula should be used among the years of observation for a given crash type. There are several existed graphical techniques for this purpose, which used to fit the crash count by type without a need to the explanatory variables (See, Mothafer et al., 2016). Later, we implemented these results to

estimate the join probability of the designated crash types among the years of observations. Followed by more investigation on the variance covariance structure and the correlation among the unobserved heterogeneity that triggered from the joint these crash types.

7.5.1 Model Specification and Crash Types Count Data Fitting

Let y_{itq} denote the observed crash count outcome of type *i* over a year of observation *t* for a given 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. Three years of observation (short balanced panel count data) is used, where *t* and index takes the value $t = 1,2,3$. We have $Q = 274$ in total segments, each segment produces three crash record reading, so we have 822 (3×274) observations in total. We assume that each crash type for a given year of observation follows a NB-II marginal distribution with a specification $F_{it}(y_{itq})$ and dispersion parameter γ_i . It's customary in crash count modeling to consider parameterizing the mean of the expected number of crashes for each crash type (denoted as λ_{itq}) as a function of all the explanatory variables $\mathbf{x}_{itq}$ with the corresponding parameters $\boldsymbol{\beta}_i$. Identifying the most significant explanatory variables vector $\mathbf{x}_{itq}$ for each crash type is required as each crash type has its own distinct mechanisms and characteristics. In this context, we selected same vectors that are represented in our previous work (see Mothafer et al., 2016) since same dataset includes crash types and explanatory variables are investigated (Ten explanatory variables were selected for the four crash count types). For the MVPCC model, we investigated (beside the graphical techniques), each of the following well-known parametric copula. Gaussian, Frank, Clayton, Gumbel and Joe copula are implemented to fit our crash types (including the explanatory variables) among the years of observations, but only the best copula in a statistical point of view will be reported to conserve on space, as we will see later.

As we mentioned before, to construct a conceivable solution for both models we need to examine the crash types in form of pairs. As for the

MVPPGC model, there are $3\times\left[\left(\frac{4-1}{2}\right)\times 4\right]=18$ pairs among the crash types and years of observation in total, while there are $4\times\left[\left(\frac{3-1}{2}\right)\times 3\right]=12$ pairs among years of observations plus (18) pairs come from joint the probability among crash type for a given year which in total (30) pairs for the MVPCC model. The copula function facilitates the correlation between the pairs of the marginal distributions of the crash types using the CML pairwise method.

7.5.1.1 Empirical Copula Diagnosis

We used the both the PP-Copula-plot and the tail-dependence graphical techniques to investigate which copula should be used among years of observation for the MVPCC model only (more details on these techniques are presented in Mothafer et al., 2016). These techniques are used to for given a preliminary idea on which copula should be used for the (12) pairs among years of observations that mentioned above. In the PP-Copula-plot, the empirical copula probability versus the theoretical parametric probability are repeated for each parametric copula that we prepared from two different copula families (elliptical vs Archimedean) and to assign the best one to our MVPCC model later. Figure (7-2) shows the several PP-Copula-plots for the same pair (rear-end crash count in 2005 vs rear end in 2006), we reported only one pair to conserve on space. With several competing parametric copulas, we prefer the one that is closest to the empirical in some sense. We can see that Gaussian; Frank and Gumbel are a good start for this pair. Other technique we used is the tail-dependence plot, the tail dependence of the rear-end in 2005 and the rear end in 2006 crashes is shown in

Figure (7-3). We can see that most of the observations given by the empirical copula are located in the upper tail (segments with higher number of crash count (more dangerous segments)) with a pattern almost similar to Frank; Gaussian copula. The same results can be deduced from other pairs (other pairs are reported Appendix.D).

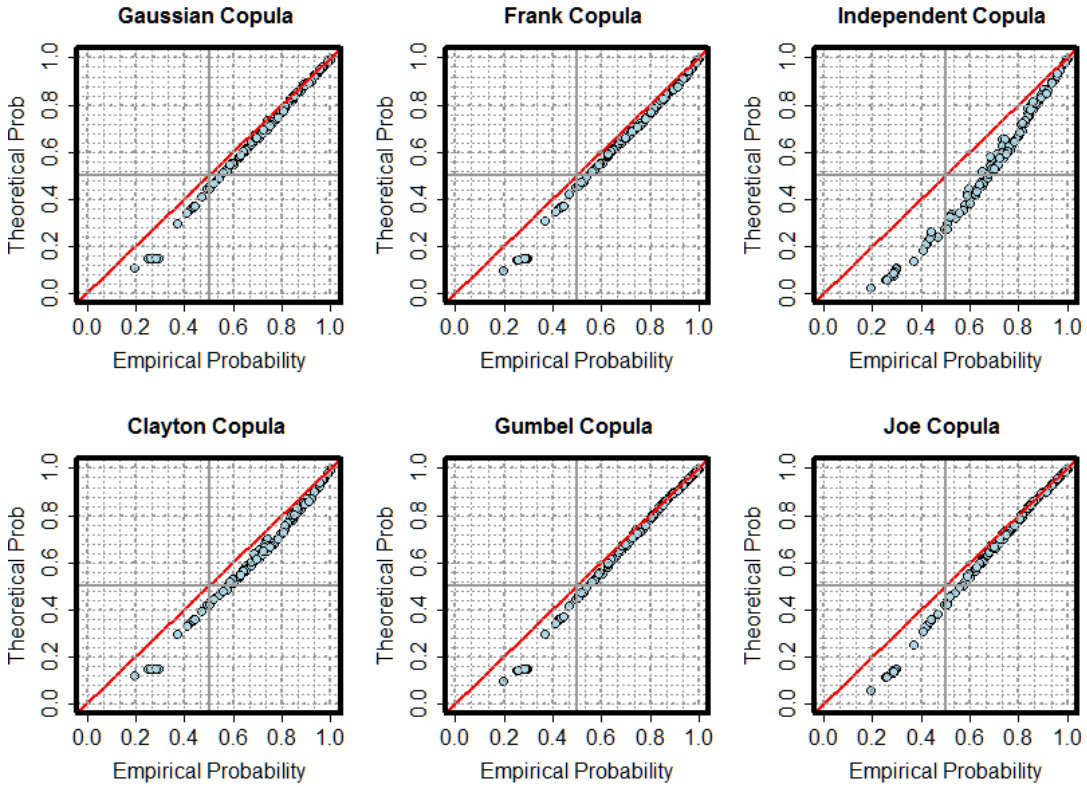


Figure (7-3) Tail-dependence plot of different copula functions for the pair rear-end 2005 against rear-end 2006

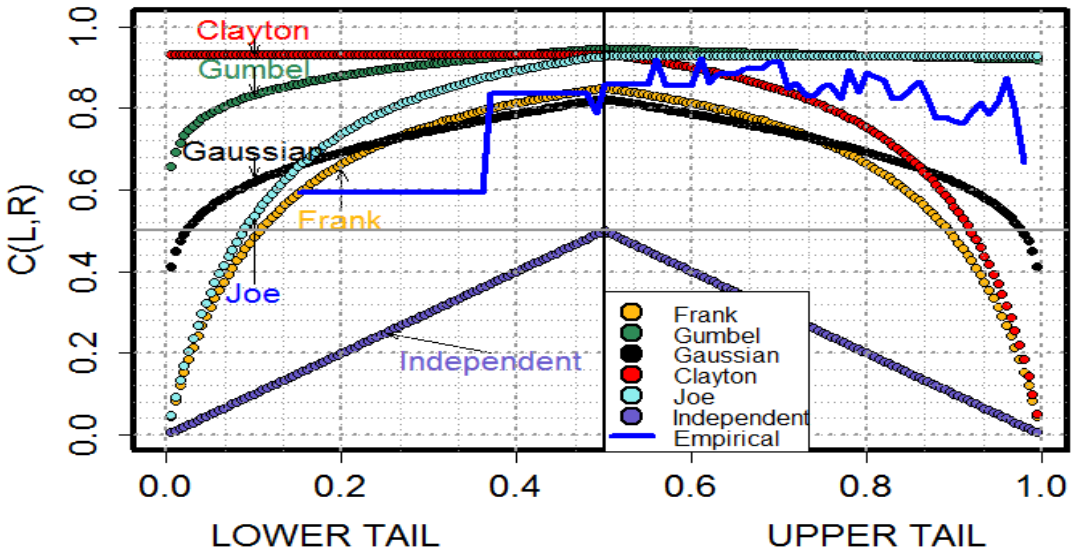


Figure (7-2) Bivariate P-P plot of different copula functions for the pair rear-end 2005 against rear-end 2006

7.5.2 Model Performance and Comparison

We have selected the *AIC*, *BIC* and the *Log likelihood* values to determine which parametric copula function is more suitable to fit the temporal effect among the years of observation for a given crash type for the MVPCC model. Table (7-1) represents the performance measure of the log-likelihood; *AIC* and *BIC* of each copula function. It is obvious that Frank copula is the most suitable to fit the temporal correlation with the highest value of log-likelihood and lower values of *AIC* and *BIC* for the MVPCC model respectively. This result is spontaneous and in concordance with the one that we obtained from the graphical techniques and now we are positively confident that Frank copula is our final choice.

A non-nested statistical likelihood ratio test was used to examine the performance of the developed models. The difference in the adjusted rho square ($\bar{\rho}_i^2$) value between the MVPPGC (the base model) and the MVPCC model (the compared model) is $4.57E-05$. The probability that this difference could have occurred by chance is equal to 0.3989 which is larger than the critical probability $\Phi\left(-\left[-2\times\left(4.57E-05\right)\times LL(C)+\left(43-39\right)\right]^{0.5}\right)$. The critical value, with a $LL(C)=-6706.81$, is equal to 0.2167 which indicates that the difference in $\bar{\rho}_i^2$ between the two discriminated models is statistically significant. In that case, MVPCC (Frank-Frank) model is more suitable to fit our crash types count data than MVPPGC, since the former offers more flexible way to model the intertemporal correlation among the years of observations for a given crash type.

The essence of the difference between the MVPPGC and the MVPCC model arises in how to model the intertemporal correlation. To make it more clear, we have isolated the components of the probabilities and evaluated the Poisson-Gamma function against the bivariate Frank copula function by fitting the crash observations without including the explanatory variables and compared them graphically (using the bivariate quantile-quantile plot). The comparison is carried out through introduce a one pair bivariate Poisson-Gamma mixture function to make it in similar to the dimensions of Frank copula

function (we used the pair rear-end crash type in 2005 with the one in 2006). There is a well-agreement between these two functions as it is shown in Figure (7-4).

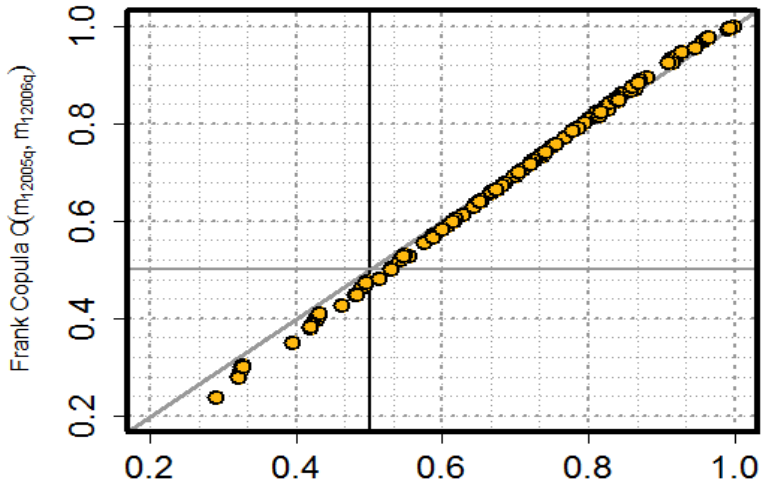


Figure (7-4) Bivariate quantile-quantile plot of bivariate frank copula vs bivariate cumulative Poisson-gamma mixture function, between observed rear-end crash in 2005 vs 2006.

Table (7-1) Log-Likelihood, Akaike Information Criterion and Bayesian Information Criterion for Various Copulas.

Copula Type	Log-Likelihood	AIC	BIC
Gaussian	-5,949	11,915	11,946
Frank	-5,898	11,812	11,843
Clayton	-6,442	12,901	12,932
Gamble	-5,901	11,820	11,851
Joe	-5,906	11,830	11,861

We also utilized a well-known standard metric measure - percentage relative error (PRE %) - to investigate the efficiency of the parameter estimates. For both models, we assumed the explanatory effects are identical and consistent across the multiple time periods. The natural comparison between the MVPPGC and the MVPCC model of the parameter estimates indicates an augmentation in efficiency. The efficiency occurred to half of the parameter

estimates of the explanatory variable vectors in the MVPCC model, due to an introduction of a more flexible function to accommodate the temporal correlation. The PRE-values of the efficient parameter estimates include the constants of each crash type mean function are ranged within [+22/-37] %. In the same context, we also took the chance to compare our superior developed model (MVPCC) against the corresponding multivariate cross-sectional model (MCORC-Frank) that we already have developed in our previous work (more details on this model can be found in Mothafer et al., 2017). We noticed immediately that the standard error of the parameters in the MCORC-Frank are significantly downward biased compared to the MVPCC model as we expected. The PRE-values between these two models among the explanatory variable vectors are ranged between [+84/-46] %. In this framework, we obtained almost more than 76% enhancement in efficiency among the parameter estimate vectors, indicates that the MVPCC model outperforms the MCORC-Frank model.

7.5.3 Empirical Estimation Results

The MVPPGC and the MVPCC model results are shown in both Table (7-2**)** and Table (7-3**)**. We discus now the parameters of the superior model only. As shown in Table (7-3**)**, the logarithm of ADT and logarithm of length are positively signed, while, the urban-rural indicator, diamond interchange indicator, shortest vertical curve length and largest rate of vertical curvature are negatively signed. Cross sections with greater than 3 lanes, number of horizontal curves and gradient of shortest vertical curve in segment are positively signed. These results obviously were similar for the MVPPGC model. The dispersion parameter γi is estimated to be 2.140, 6.167, 4.821 and 7.016 for rear-end, sideswipe, fixed object and all-other types respectively, which imply an overdispersion magnitude of 0.467, 0.162, 0.207 and 0.143. The small size effect of the time-invariant of the unobserved heterogeneity for the “all-others” type indicates that it’s possible to use Poisson marginal distribution, which matches the MCORC-Frank finding. The intertemporal correlation parameters ϑi are positively signed with values 2.140, 6.167, 4.821 and 7.016 for rear-end, sideswipe, fixed object and all-other types respectively. As its

evident from above, our superior model (the MVPCC) supports different sizes of overdispersion and serial correlation in order to represent both the time-invariant and time-varying correlations of the unobserved heterogeneity for each frequency of our crash type. The interpretation of the parameter estimates is in agreement with our previous work in Chapter 4Chapter 5.

7.5.4 Variance-Covariance Representation

The parameters of the level of dependency θ_{ij} are presented also in same previous Table (7-3**)**. These parameters demonstrate the presence of common unobserved factors among the crash types assuming that the latter will not vary among the years of observations. A significant positive value of θij indicates an association between the unobserved factors of each crash type in the corresponding pair. Rear-end vs sideswipe, rear-end vs all-others, sideswipe vs fixed objects; sideswipe vs all-others and finally fixed object vs all-others are found to be statistically significant, except the rear-end vs fixed object pair which found not significant indicates no correlation between these two crash types.

The varaince-covaraince matrices, of the MVPCC model of the dynamic unobserved heterogeneity were estimated considering the average values among all segments using both Eq. (7-14) and its equal to,

$$\mathbf{VT}_{1\,(3,3)} = \begin{bmatrix} 78.52 & & \\ 154.39 & 91.20 & \\ 167.00 & 180.66 & 105.97 \end{bmatrix}_{i=1}, \mathbf{VT}_{2\,(3,3)} = \begin{bmatrix} 4.90 & & \\ 10.06 & 5.29 & \\ 10.60 & 11.18 & 5.71 \end{bmatrix}_{i=2},$$
$$\mathbf{VT}_{3\,(3,3)} = \begin{bmatrix} 3.98 & & \\ 4.00 & 4.15 & \\ 4.13 & 4.25 & 4.33 \end{bmatrix}_{i=3}, \mathbf{VT}_{4\,(3,3)} = \begin{bmatrix} 2.25 & & \\ 1.65 & 2.35 & \\ 1.72 & 1.78 & 2.45 \end{bmatrix}_{i=4} \quad (7\text{-}25)$$

The results in Eq.(7-25) indicate the differences in size of the variance-covariance among the years of observations for a given crash type. Since the rear end crash type was the highest crash record in our dataset, it was natural to see that $\mathbf{VT}_1$ was the highest values. The covariance among the crash type pairs for the MVPCC model for a given year of observation is given as,

Table (7-2) Multivariate panel Poisson gamma mixture -copula based model: Frank Copula

Explanatory variables	Rear-End		Sideswipe		Fixed Objects		All-Others	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Constant	-23.497	2.367	-16.175	1.507	-7.527	2.134	-10.059	1.067
LnAADT	2.493	0.244	1.688	0.155	0.847	0.211	1.064	0.112
LnLength	0.486	0.160	0.657	0.118	0.844	0.128	0.952	0.080
Urban rural dummy, 1 if rural,0 if urban	-0.620	0.201	-0.662	0.146	-0.137♣	0.145		
Proportion of three or more lanes cross section by length of segment	0.897	0.135	0.610	0.118	0.348	0.110	0.476	0.107
Number of horizontal curves per segment	0.149	0.058	0.101	0.037	0.055‡	0.037		
Diamond interchange type dummy	-0.207†	0.122	-0.266	0.094	-0.204†	0.111		
Smallest vertical gradient in segment			0.068*	0.033				
Largest beginning vertical curve elevation in segment			0.710†	0.389				
Largest horizontal curve central angle in segment					0.555*	0.255		
Number of vertical curves in segment					-0.044♣	0.036		
Dispersion parameter γi (intertemporal effect is included)	1.946*	0.203	5.420	0.909	4.398	0.901	6.485	1.500
Level of Dependency θij								
Sideswipe	2.058	0.337						
Fixed Objects	0.403♣	0.316	1.125	0.284				
All-others	0.655*	0.255	0.985	0.285	1.171	0.265		
Sample size	273							
LL (C) with constant parameters only	-6,706.8							
LL (β) at convergence	-5,928.3							
AIC	11876.1							
BIC	11911.3							

S.E. standard error; - Not relevant; * Significant at 5% level; † Significant at 10% level; ‡ Significant at 15% level; ♣ Not Significant. All the other coefficients are significant at the level of 1%. Significance of the actual overdispersion parameter ($1/\gamma_i$) is estimated using the delta method. Significance of the overdispersion effect is very strong for all crash types, at or better than the 99.5% level except the one for rear-end crash type.

Table (7-3) Multivariate panel copula -copula based model: Frank-Frank copula

Explanatory variables	Rear-End		Sideswipe		Fixed Objects		All-Others	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Constant	-24.225	2.443	-15.894	1.596	-8.465	1.719	-10.044	1.055
LnAADT	2.564	0.250	1.658	0.165	0.935	0.174	1.061	0.111
LnLength	0.456	0.166	0.640	0.122	0.809	0.113	0.944	0.081
Urban rural dummy, 1 if rural,0 if urban	-0.563	0.214	-0.657	0.143	-0.086♣	0.122		
Proportion of three or more lanes cross section by length of segment	0.852	0.136	0.610	0.121	0.346	0.106	0.478	0.107
Number of horizontal curves per segment	0.140*	0.058	0.096*	0.038	0.047‡	0.032		
Diamond interchange type dummy	-0.163♣	0.156	-0.271	0.096	-0.169†	0.096		
Smallest vertical gradient in segment			0.070*	0.032				
Largest beginning vertical curve elevation in segment			0.742†	0.418				
Largest horizontal curve central angle in segment					0.643	0.245		
Number of vertical curves in segment					-0.031♣	0.026		
Dispersion parameter γi	2.140	0.247	6.167	1.314	4.821	0.998	7.016	1.924
Intertemporal Correlation parameter ϑi	4.046	0.487	2.208	0.365	1.613	0.299	0.994	0.262
Level of Dependency θij (among crash types)								
Sideswipe	1.939	0.334						
Fixed Objects	0.356♣	0.258	1.116	0.279				
All-others	0.643	0.232	0.968	0.284	1.165	0.258		
Sample size	273							
LL (C) with constant parameters only	-6,677.1							
LL (β) at convergence	-5,897.5							
AIC	11,812.2							
BIC	11,843.3							

S.E. standard error; - Not relevant; * Significant at 5% level; † Significant at 10% level; ‡ Significant at 15% level; ♣ Not Significant. All the other coefficients are significant at the level of 1%.

$$\mathbf{VI}_{(4,4)1} = \begin{bmatrix} 78.52 & & & \\ 64.48 & 4.90 & & \\ 32.72 & 10.47 & 3.98 & \\ 22.31 & 6.95 & 4.80 & 2.25 \end{bmatrix}_{2005},$$

$$\mathbf{VI}_{(4,4)2} = \begin{bmatrix} 91.20 & & & \\ 73.47 & 5.29 & & \\ 36.48 & 11.37 & 4.15 & \\ 24.99 & 7.58 & 5.12 & 2.35 \end{bmatrix}_{2006}, \quad (7\text{-}26)$$

$$\mathbf{VI}_{(4,4)3} = \begin{bmatrix} 105.97 & & & \\ 83.70 & 5.71 & & \\ 40.67 & 12.34 & 4.32 & \\ 27.99 & 8.27 & 5.46 & 2.45 \end{bmatrix}_{2007}$$

The average of the covariance values among the years of observations given in Eq.(7-26) concurs the results of the MCORC-Frank finding since both the developed model and the former model utilize the Frank copula among the crash types (see Mothafer et al., 2016). The total covariance values of both the MVPPGC and the MVPCC model of the ones resulting from estimates the expected number of crashes of a specific type and the ones from the stochastic error term associations generated from each marginal pair among the years of observation are given in Table (7-4). The negative value of the PRE% represents the amount of the crashes that unexplained by the designated model, as this percentage decreases, the model is less statistically preferable. The average of PRE% values for the MVPCC model (-16%) which is larger than the corresponding average of the MVPPGC model (-36%). The results suggest that the MVPCC model once again outperforms the MVPPGC in representing more accurately the covariance structure among the years of observations. In an analogous pattern, we also computed the covariances among the crash types for a given year of observation and presented them in Table (7-5)[13]. As it evident through this table, our proposed model offers more

[13] Theoretically speaking, the average of these values among the years of observations are similar to the one we obtained using the MCORC-Frank model. For example, the average value of the covariance among the years of observations between the rear end and the sideswipe for our superior model is 73.88 while for our previous model it was 124.12 [Table (5-7) in Mothafer et al.,2016]

details compared to our cross-sectional previous model. As it evident through this table, our proposed model offers more details compared to our cross-sectional previous model.

The last step is to calculate the total variance structure among the years of observations for a given crash type for both models since they used two different concepts to model the intertemporal correlations. The results are presented in Table (7-6), which indicate that the MVPCC model has an average of PRE% (-35%) which is larger than the corresponding value of the MVPGC model (-45%). These results reflect which model is more accurately performed better in predicating the crash count. Table (7-7) epitomizes the total variance values among crash types for a given year of observation of our superior model, which are compared to the corresponding observed ones. The average of these values is almost similar to the one we obtained before in [Table (5-8)].

Table (7-4) Total covariance among the years of observations for a given crash type.

	Rear-End			Sideswipe			Fixed Objects		
	Obs.	MVPCC (Frank-Frank)		Obs.	MVPCC (Frank-Frank)		Obs.	MVPCC (Frank-Frank)	
		Total Cov	PRE%		Total Cov	PRE%		Total Cov	PRE%
2005									
Sideswipe	109.09	64.48	-41%						
Fixed Objects	54.65	32.72	-40%	15.62	10.47	-33%			
All Others	31.91	22.31	-30%	9.4	6.95	-26%	6.46	4.8	-26%
2006									
Sideswipe	105.02	73.47	-30%						
Fixed Objects	47.92	36.48	-24%	14.24	11.37	-20%			
All Others	33.62	24.99	-26%	10.52	7.58	-28%	6.17	5.12	-17%
2007									
Sideswipe	98.99	83.7	-15%						
Fixed Objects	36.05	40.67	13%	11.45	12.34	8%			
All Others	22.12	27.99	27%	6.82	8.27	21%	3.29	5.46	66%

PRE %= (Estimated-Obs.)/Obs. 100%

Table (7-5) Total covariance among the crash types for a given year of observation for the MVPCC model.

	Rear-End			Sideswipe			Fixed Objects		
	Obs.	MVPCC (Frank-Frank)		Obs.	MVPCC (Frank-Frank)		Obs.	MVPCC (Frank-Frank)	
		Total Cov	PRE%		Total Cov	PRE%		Total Cov	PRE%
2005									
Sideswipe	109.09	64.48	-41%						
Fixed Objects	54.65	32.72	-40%	15.62	10.47	-33%			
All Others	31.91	22.31	-30%	9.4	6.95	-26%	6.46	4.8	-26%
2006									
Sideswipe	105.02	73.47	-30%						
Fixed Objects	47.92	36.48	-24%	14.24	11.37	-20%			
All Others	33.62	24.99	-26%	10.52	7.58	-28%	6.17	5.12	-17%
2007									
Sideswipe	98.99	83.7	-15%						
Fixed Objects	36.05	40.67	13%	11.45	12.34	8%			
All Others	22.12	27.99	27%	6.82	8.27	21%	3.29	5.46	66%

PRE %= (Estimated-Obs.)/Obs. 100%

Table (7-6) Total variance structure among the years of observations for a given crash type

		MVPPGC Model				MVPCC Model (Frank-Frank)			
Crash Type	Observed	$V_i[E(y_{it}^{Time})]$	$E[V_i(y_{it}^{Time})]$	$V_i^{Time}(y_{it})$	PRE%	$V_i[E(y_{it}^{Time})]$	$E[V_i(y_{it}^{Time})]$	$V_i^{Time}(y_{it})$	PRE%
Rear End	3651.11	1347.16	993.36	2340.52	-36%	1143.30	1279.81	2423.12	-34%
Sideswipe	277.47	68.13	38.08	106.21	-62%	61.94	79.59	141.52	-49%
Fixed Object	82.44	23.02	25.59	48.60	-41%	23.05	37.22	60.27	-27%
All-Others	40.78	11.62	11.72	23.34	-43%	11.30	17.34	28.64	-30%

Table (7-7) Total variance among crash types for a given year of observation.

		MVPCC (Frank-Frank)			
Year	Observed	$V_t[E(y_{it}^{Crash})]$	$E[V_t(y_{it}^{Crash})]$	$V_t^{Crash}(y_{it})$	PRE%
2005	985.02	300.77	272.21	572.98	-42%
2006	901.74	341.56	308.37	649.94	-28%
2007	764.92	388.15	349.55	737.70	-4%

7.6 SUMMARY

This chapter proposes an econometric framework to model the multivariate panel crash by type count data. The point of interest in the multivariate panel modeling is to obtain more efficient parameter estimates by taking in consideration the intertemporal correlations of a given crash type among the years of observations. Formulating the joint probability among different crash types also triggers an inter-type correlations which added more difficult to seek a conceivable inference. Our effort was to simplify this problem by introduce a flexible solution (computationally tractable), that is in no need for a simulation-based technique, which is a common case in modeling multivariate panel data. Two proposed models are introduced for this purpose, the MVPGC and the MVPCC model. In first model, we used the random effect principle to model the intertemporal correlation among the years of observations for a given crash type. Poisson-Gamma Mixture function is used as an easy distribution to imply in that context, while the bivariate copula function was exploited in the second model for same correlation. Both models use the bivariate copula function to model the inter-type correlations (among crash types) for a given year of observation. We sought a solution for the joint probability of these two models through the pairwise copula-CML technique which offers a joint distribution without any restrictions to accept both positive and negative correlations.

The performance of these proposed models is demonstrated through an empirical application to study four different categories of crash types that commonly occur on freeway segments located on highway No. 5 in the State of Washington, USA for three years of observations (from 2005 to 2007, balanced short panel). The statistical superior model is used to draw an inference on the dependence structure among these categories. The empirical results show that Frank copula is more preferable to fit the intertemporal correlations which allows more freedom to the unobserved heterogeneity to interact, compared to the Poisson-Gamma mixture distribution. As we expected, the standard errors of the estimated parameters are more efficient if it's compared to the corresponding downward biased parameters of the multivariate cross-sectional count model. The variance-covariance structure is more accurately represented by our proposed model with more ability in

predicating the crash count outcomes.

The severity level is not considered in this paper. One might think of utilizing another crash count dataset that contains the number of crashes by both crash type and severity level which offers richer insights into the differential impacts of various explanatory variables on the crashes.

Chapter 8

Conclusions and Future Work

8.1 CONCLUSIONS

In the context of this thesis, we have developed several multivariate count models to accommodate the correlation among different crash types. The point of interest in the multivariate modeling are to investigate whether these crash types are jointly determined and to enhance the parameter estimate efficiency. The unobserved heterogeneity (factors) associations with crash-count data are manifested into three forms mainly, which are: A) the ones causes the over-dispersion in the context of cross-sectional data (or within a short observation period like one year in the panel data context). B) The one arises from the correlation of the error term of same crash type outcome over the same observation unit correlated among the years of observations (also known as serial correlation, autocorrelation and intertemporal effect in the literature). C) The ones which are triggered from the association of the unobserved heterogeneity among different types of crashes (inter-type correlation). In our consideration, we set a framework that serve our intention to produce multivariate count model with non-biased parameter estimates. Furthermore, we sought a more practical methodological solution that can be less computational expensive and more accurately demonstrate crash-count predictions. Our crash data record are obtained from the Washington transportation department. The crash-count by type observations consist a sample size produces 822 (=274x3) segment-year observations. rear-end, sideswipe, and fixed objects crash types and “all other,” types are considered in each chapter in this study.

Here and after, we will briefly confer an overview of the main findings of

each chapter in this thesis, so that, next section will discuss our contributions to the crash-count modeling followed by a section on the points that to be considered in the future researches.

8.2 CONTRIBUTIONS

In chapter 4, we have developed the multivariate Poisson gamma mixture model (MVPGM).The crash-data record treated as a cross-sectional data, thus no time is involved in calculation. The model is used to investigate the inter-type correlations and covariances structure among the designated crash types. This model has a closed-form for the joint probability function and ease to implement but with so called a 'single-shared' property. This property has been assumed to compensate the intricacy of the joint probability on expensive that all the unobserved heterogeneity from each different crash type carries same size. Moreover, MVPGM is uniquely restricted to represent only the positive correlation among crash types. The model parameters show that indeed there are significant unobserved heterogeneity correlations. In addition the results show that MVPGM covariances of crash types are in better agreement with observed covariances than those from univariate crash type models. These findings are in spite of our observation that the individual crash type models provide for statically better fits due to their unconstrained dispersion parameters, which is constrained in our proposed model here. This finding underscores the need to explore the behavior of dispersion in multivariate crash type contexts.

Our next task was to extend the work of chapter four, through seeking a better crash-count model under same assumption of cross-sectional data scope. Thus, we developed the multivariate copula-based ordered response model (MCROC) in chapter five. This model promises a joint probability without any restrictions to allow both positive/negative correlations among the error terms structure components. Likewise, more tractable model that is in no need for a simulation-based solution which is a common computationally burdensome solution that appears in many multivariate count models. The model utilizes an alternative way to approach the problem which is represented through modifying the conventional ordered response model. This modification

exploits the latent continuous variable to match the corresponding count variable through considering their probability equivalency concept. Therefore to serve all our preferences above we have employed both the CML technique and the bivariate copula function that both show an excellent cooperation over offering a parametric, straightforward and flexible model. In addition to the above features, our proposed model adds a plus-point over representing more accurately the correlation structure through allowing a non-linear asymmetric schemes among the unobserved heterogeneity types. The empirical results show that Frank copula is the best to fit our crash-count data from the statistical point of view. Furthermore, considering the correlations among the crash types are highly beneficial and recommended to enhance the covariance and total variance structure that both present a better prediction.

Be in quest of incorporate the period of analysis (time of observation) effect among the crash types which is neglected in both chapter four and chapter five, we presented our effort in this context in chapter six. The crash-count data are now considered as a panel-count data were the time of observation is relevant. We evaluated the random effects Poisson gamma model (REPG) against a crash sum model for three years of observations. Our major findings here was that the REPG model and crash sum model were empirically equivalent based on several statistical tests of parameter estimates similarities. The crash sum model appears to be a reasonable empirical alternative to the panel model given that the standard errors and parameter magnitudes are highly similar in the presence of time invariant heterogeneity. The reasonableness of the crash sum model as an alternative appeals from the fact that it requires less frequent measurement of geometrics and ADT. This is under the assumption that dramatic changes in the independent variable vector do not occur in the panel observation period. Although, our major concern was whether this assumption is valid, the capture of time invariant heterogeneity via the crash sum model ensures that the appropriate variables are identified as statically significant in the model. We recommend to use the crash sum model as an alternative to the EPRG model with for short panel crash-count data only an extra caution.

Through the transition from cross-sectional crash-count data modeling

presented in chapter four and five, into panel crash count modeling presented in chapter six, we maintained the same focus but to expand the work of chapter six to develop our last model in this thesis. Chapter seven, proposes an econometric scheme to model the multivariate panel crash-count by type data. The work here is an effort to combine and promote all the techniques that we learned in chapter four, five and six. To our basic knowledge, our effort in this chapter is considered as a first attempt to model multivariate panel crash-count data. We sought a conceivable inference through developing a flexible, computationally tractable model that takes in consideration both the time invariant and time-varying unobserved heterogeneity effects. For this purpose, we proposed two models, the multivariate panel Poisson gamma-copula model (MVPPGC) and the multivariate panel copula-copula model (MVPCC). In the first model we utilized from chapter six the random effect Poisson gamma function to accommodate the correlation among the years of observations, while the bivariate copula with the pairwise CML are used to carry the correlations among the crash types. In the second model we replaced the Poisson gamma function by the bivariate copula and pairwise CML to achieve the same goal. The empirical results show that Frank copula is more preferable to fit the time-invariant unobserved heterogeneity among different crash types across the years of observations. It was not a surprise that our superior proposed model (MVPCC) produces a more efficient parameter of estimates through enlarging the standard error, if they are compared to the corresponding downward biased parameters of the multivariate cross-sectional count model that presented in chapter five. Thus, the Table (7-3) represents our last update for the parameter estimates which can offer a valuable insight on each crash count type related parameters. Since only the standard errors are different between MVPGM and MVPCC we can refer to same extensive discussion on the importance of these explanatory variables for each crash count type chapter four with consideration our final results here.

8.3 FUTURE RESEARCH

Within the objectives of this study, we have overcome several methodological crash-count related problems. Nevertheless, we haven't included other

perspectives which offer a richer insight regarding crash count by type. For example, the severity level is not considered due to our limited current crash data information. Perhaps, crash count by type categorized by severity level will serve this perspective. Furthermore, increase the sample size and analyze a larger panel count data will enhance our parameter estimate standard errors efficiency even further than the current status. We believe that the proposed methodology framework in this study, can be transferred easily to other data set. Not to mention that we can add more dimensions of several crash count outcomes (for example disassembly the all-other category crash types into other dimensions) to investigate the usability of our current model in regard or time/computational cost. The interesting part of the empirical copula that we presented among different pairs in our crash count dataset is that we can use different copula function for each pair based on the graphical techniques that we demonstrated. In this regard, further work is possible to get more efficient model, rather than our current assumption, that copula function type is a fixed type among all the pairs.

Bibliography

Abdelwahab, H.T., Abdel-Aty, M.A., 2002. Artificial neural networks and logit models for traffic safety analysis of toll plazas. Transportation Research Record 1784, 115–125.

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.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55–63.

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

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.

Aptech, 1998. Gauss 3.5 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.

Baltagi, B. (2008). *Econometric Analysis of Panel Data*. John Wiley and Sons.

Ben-Akiva, M.E., and Lerman, S.R. (1985). *Discrete Choice Analysis: Theory and application to travel demand*. MIT Press.

Bhargava, A., Franzini, L. and Narendranathan, W., 1982. Serial correlation and the fixed effects model. The Review of Economic Studies, 49(4), pp.533-549.

Bhat, C. R., Born, K., Sidharthan, R., and Bhat, P. C. (2014). 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. and 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), pp.243-272.

Bhat, C.R., 2011. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. Transportation Research Part B: Methodological, 45(7), pp.923-939.

Bhat, C.R., 2014. The Composite Marginal Likelihood (CML) Inference Approach with Applications to Discrete and Mixed Dependent Variable Models. Foundations and Trends® in Econometrics, 7(1), pp.1-117.

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., Eluru, N., 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. Transportation Research Part B: Methodological, 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, Volume 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.

Bijleveld, F.D., 2005. The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. Accident Analysis and Prevention 37 (4), 591–600.

Bonneson, J.A., McCoy, P., 1993. Estimation of safety at two-way stop-controlled intersections on rural roads. Transportation Research Record 1401, 83–89.

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), pp.135-162.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., Roncalli, T., 2000. Copulas for finance-a reading guide and some applications. Available at SSRN 1032533, 69 pages.

Brüde, U., Larsson, J., 1993. Models for predicting accidents at junctions where pedestrians and cyclists are involved. How well do they fit? Accident Analysis and Prevention 25 (5), 499–509.

Cameron, A.C. and Trivedi, P.K. (2001). Essentials of count data regression. (In B. H. Baltagi.) *A Companion to Theoretical Econometrics*, 331-348.

Cameron, A.C. and Trivedi, P.K. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Cameron, A.C. and Trivedi, P.K., 2013. Regression analysis of count data (Vol. 53). Cambridge university press.

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, C., Trivedi, P., (1986). Econometric models based on count data: comparisons and applications of some estimators and tests. *Journal of Applied Econometrics* 1(1), 29-53.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on accident frequencies and severities. Accident Analysis and Prevention 33 (1), 99–109.

Castro, M., Paleti, R., and 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*: *Methodological*, 46(1), 253–272.

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

Chang, H.-L., Jovanis, P.P., 1990. Formulating accident occurrence as a survival process. Accident Analysis and Prevention 22 (5), 407–419.

Chang, L.-Y., 2005. Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. Safety Science 43 (8), 541–557.

Chin, H.C. and Quddus, M.A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis and Prevention*, 35(2), 253-259.

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

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.

Daniels, S., Brijs, T., Nuyts, E., Wets, G., 2010. Explaining variation in safety performance of roundabouts. Accident Analysis and Prevention.

Das, A., and Abdel-Aty, M. A. (2011). A combined frequency–severity approach for the analysis of rear-end crashes on urban arterials. *Safety Science*, 49(8-9), 1156–1163.

Deheuvels, P., 1979. La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance. Acad. Roy. Belg. Bull. Cl. Sci. (5), 65(6), pp.274-292.

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

Dey, D.K. and Y. Chung (1992) Compound Poisson distributions: properties and estimation, *Communications in Statistics – Theory and Methods* 21: 3097-3121.

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., and 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. and 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: methodological, 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. and Chakraborti, S., 2011. Nonparametric statistical inference (pp. 977-979). Springer Berlin Heidelberg.

Gkritza, K., Kinzenbaw, C. R., Hallmark, S. and Hawkins, N. (2010). An empirical analysis of farm vehicle crash injury severities on Iowa's public road system. *Accident Analysis and Prevention*, 42(4), 1392–1397.

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

Green, W.H. (2003). *Econometric Analysis.* Pearson Education India.

Gurmu, S. and Elder, J. (2000) Generalized bivariate count data regression models, *Economics Letters*, 68, 31–36.

Halekoh, U., Højsgaard, S., Yan, J., 2006. The R Package geepack for generalized estimating equations. Journal of Statistical Software 15 (2), 1–11.

Hauer, E., Ng, J.C.N., Lovell, J., 1988. Estimation of safety at signalized intersections. Transportation Research Record 1185, 48–61.

Hausman, J., B.H. Hall and Z. Griliches (1984). Econometric models for count data with an application to the Patents-R, and D relationship, *Econometrica* 52: 909-938.

Hernández-Maldonado, V., Díaz-Viera, M. and 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, pp.112-123.

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 & Probability Letters, 7(4), 283-286.

Imprialou, M.I.M., Quddus, M. and Pitfield, D.E., 2016. Predicting the safety impact of a speed limit increase using condition-based multivariate Poisson lognormal regression. Transportation Planning and Technology, 39(1), pp.3-23.

Ivan, J.N., Wang, C. and Bernardo, N.R. (2000). Explaining two-lane highway crash rates using land use and hourly exposure. *Accident Analysis and Prevention*, 32(6), 787-795.

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.

Johansson, P., 1996. Speed limitation and motorway casualties: a time series count data regression approach. Accident Analysis and Prevention 28 (1), 73–87.

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accident Analysis and Prevention 35 (1), 59–69.

Jones, B., Janssen, L., Mannering, F., 1991. Analysis of the frequency and duration of freeway accidents in Seattle. Accident Analysis and Prevention 23 (2), 239–255.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transportation Planning and Technology 15 (1), 41–58.

Jovanis, P.P., Chang, H.L., 1986. Modeling the relationship of accidents to miles traveled. Transportation Research Record 1068, 42–51.

Jovanis, P.P., Chang, H.L., 1989. Disaggregate model of highway accident occurrence using survival theory. Accident Analysis and Prevention 21 (5), 445–458.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F. and Jöreskog, K.G., 2012. Pairwise likelihood estimation for factor analysis models with ordinal data. Computational Statistics & Data Analysis, 56(12), pp.4243-4258.

Katuwandeniyage, H. and Priyantha, K., 2015. Multivariate longitudinal data analysis for actuarial applications.

Kim, D.-G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis and Prevention 39 (1), 125–134.

Kim, D.G., Washington, S. and Oh, J., 2006. Modeling crash types: New insights into the effects of covariates on crashes at rural intersections. Journal of Transportation Engineering, 132(4), pp.282-292.

Kockelman K. M. (2001). A model for time- and budget- constrained activity demand analysis, *Transportation Research Part B*,35, 225-269.

Kuha, J. (2004). AIC and BIC comparisons of assumptions and performance. *Sociological Methods and Research*, 33(2), 188-229.

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

Kumara, S. and Chin, H. (2004). Study of fatal traffic accidents in Asia Pacific countries. *Transportation Research Record: Journal of the Transportation Research Board*, 1897, 43-47.

Kumara, S.S.P., Chin, H.C., 2003. Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros. Traffic Injury Prevention 3 (4), 53–57.

Law, T.H., Noland, R.B. and Evans, A.W. (2009). Factors associated with the relationship between motorcycle deaths and economic growth. *Accident Analysis and Prevention*, 41(2), 234-240.

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. and Jiang, X., 2015. Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis & Prevention, 78, pp.146-154.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accident Analysis and Prevention 34 (2), 149–161.

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, X., Lord, D., Zhang, Y., 2009. Development of accident modification factors for rural frontage road segments in Texas using results from generalized additive models. Working Paper, Zachry Department of Civil Engineering, Texas A&M University, and College Station, TX.

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

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

Lord, D. (2000). The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models (Doctoral dissertation, University of Toronto).

Lord, D. and Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological

alternatives. *Transportation Research Part A: Policy and Practice*, 44(5), 291–305.

Lord, D., Geedipally, S.R., Guikema, S., 2010. Extension of the application of Conway–Maxwell–Poisson models: analyzing traffic crash data exhibiting underdispersion. Risk Analysis.

Lord, D., Guikema, S., Geedipally, S.R., 2008. Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention 40 (3), 1123–1134.

Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-v/c ratio for rural and urban freeway segments. Accident Analysis and Prevention 37 (1), 185–199.

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Safety Science 46 (5), 751–770.

Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transportation Research Record 1717, 102–108.

Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, Poisson-gamma and zero-inflated regression models for motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis and Prevention* 37, 35-46.

Lum, H. and Reagan, J.A., 1995. Interactive highway safety design model: accident predictive module. Public Roads, 58(3).

Lyon, C., Oh, J., Persaud, B., Washington, S. and Bared, J. (2003). Empirical investigation of interactive highway safety design model accident prediction algorithm: Rural intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 1840, 78-86.

Ma, J., and Kockelman, K.M., (2006). Bayesian multivariate Poisson regression for models of injury count by severity. *Transportation Research Record* 1950, 24-34.

Ma, J., Kockelman, K. M., and 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.

Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accident Analysis and Prevention 41 (2), 217–226.

Mannering, F. L. and Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1, 1–22.

Mannering, F., Shankar, V., Bhat, C. (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.

Maycock, G., Hall, R.D., 1984. Accidents at 4-Arm Roundabouts. TRRL Laboratory Report 1120, Transportation and Road Research Laboratory, Crow Thorne, UK.

Miaou, S.P. and Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record: Journal of the Transportation Research Board*, 1840, 31-40.

Miaou, S.P. and Song, J.J. (2005). Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention*, 37(4), 699-720.

Miaou, S.-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26 (4), 471–482.

Miaou, S.-P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25 (6), 689–709.

Miles, D. (2001). Joint purchasing decisions: A multivariate negative binomial approach. *Applied Economics*, 33, 937-946.

Mothafer, G.I., Yamamoto, T. and 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.

N'Guessan, A., Langrand, C., 2005a. A covariance components estimation procedure when modelling a road safety measure in terms of linear constraints. Statistics 39 (4), 303–314.

N'Guessan, A., Langrand, C., 2005b. A Schur complement approach for computing sub-covariance matrices arising in a road safety measure modeling. Journal of Computational and Applied Mathematics 177, 331–345.

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.

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.

Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accident Analysis and Prevention 38 (2), 346–356.

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, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683–691.

Park, E. S., and Lord, D. (2007). Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record*, 2019, 1–6.

Persaud, B.P., 1994. Accident prediction models for rural roads. Canadian Journal of Civil Engineering 21 (4), 547–554.

Quddus, M.A. (2008). Time series count data models: An empirical application to traffic accidents. *Accident Analysis and Prevention*, 40(5), 1732-1741.

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.

Rumar, K., 1985. The role of perceptual and cognitive filters in observed behavior. In Human behavior and traffic safety (pp. 151-170). Springer US.

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.

Shaheed, M. S. B., Gkritza, K., Zhang, W., and Hans, Z. (2013). A mixed logit analysis of two-vehicle crash severities involving a motorcycle. *Accident Analysis and Prevention*, 61, 119–28.

Shankar, V., Albin, R., Milton, J. and Mannering, F. (1998). Evaluating median crossover likelihoods with clustered accident counts: An empirical inquiry using the random effects negative binomial model. *Transportation Research Record: Journal of the Transportation Research Board*, 1635, 44-48.

Shankar, V., Mannering, F. and Barfield, W., (1995). Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis and Prevention*, 27(3), 371-389.

Shankar, V., Mannering, F., and Barfield, W. (1996). Statistical analysis of accident severity on rural freeways. *Accident Analysis and Prevention* 28(3), 391–401.

Shankar, V., Milton, J., Mannering, F.L., 1997. Modeling accident frequency as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention 29 (6), 829–837.

Shankar, V.N., Albin, R.B., Milton, J.C., Mannering, F.L., 1998. Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using random effects negative binomial model. Transportation Research Record 1635, 44–48.

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

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

Sittikariya, S., Shankar, V.N., Shyu, M.B. and Chayanan, S. (2005). Accounting for serial correlation in count models of traffic safety. *Journal of the Eastern Asia Society for Transportation Studies*, 6, 3645-3657.

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

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. Journal of Multivariate Analysis 97(1), 246–273.

Spissu, E., Pinjari, A.R., Pendyala, R.M. and Bhat, C.R., 2009. A copula-based joint multinomial discrete–continuous model of vehicle type choice and miles of travel. Transportation, 36(4), pp.403-422.

Stata Corp LP. (2005). *Stata Statistical Software Release 9*. Stata Press Publication.

Sun, J., Frees, E.W. and 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.

Ulfarsson, G., and Shankar, V. (2003). Accident count model based on multiyear cross-sectional roadway data with serial correlation.

Transportation Research Record: Journal of the Transportation Research Board, 1840, 193-197.

Van Ophem, H., 1999. A general method to estimate correlated discrete random variables. Econometric Theory, 15(02), 228-237.

Varin, C., 2008. On composite marginal likelihoods. AStA Advances in Statistical Analysis, 92(1), pp.1-28.

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

Winkelmann, R., 2013. Econometric analysis of count data. Springer Science & Business Media.

Winklemann, R. (2008). *Econometric Analysis of Count Data*. Springer-Verlag Berlin Heidelberg (Vol. Fifth edit, p. 349).

Xie, Y., Zhang, Y., 2008. Crash frequency analysis with generalized additive models. Transportation Research Record 2061, 39–45.

Xu, J.J., 1996. Statistical modelling and inference for multivariate and longitudinal discrete response data (Doctoral dissertation, University of British Columbia).

Yaacob, W.F.W., Lazim, M.A. and Wah, Y.B., 2012. Modeling Road Accidents using Fixed Effects Model: Conditional versus Unconditional Model. In Proceedings of the World Congress on Engineering (Vol. 1).

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.

Yan, X., Ma, M., Huang, H., Abdel-Aty, M., and Wu, C. (2011). Motor vehicle-bicycle crashes in Beijing: irregular maneuvers, crash patterns, and injury severity. *Accident Analysis and Prevention*, 43(5), 1751–8.

Yang, Z., Zhibin, L., Pan, L., and Liteng, Z. (2011). Exploring contributing factors to crash injury severity at freeway diverge areas using ordered probit model. *Procedia Engineering*, 21, 178–185.

Yasmin, S., Eluru, N., Bhat, C. R., and Tay, R. (2014). A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic Methods in Accident Research*, 1, 23–38.

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

Ye, X., Pendyala, R. M., Washington, S. P., Konduri, K., and Oh, J. (2009). A simultaneous equations model of crash frequency by collision type for rural intersections. *Safety Science*, 47(3), 443–452.

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

Appendices

Appendix.A

Cross-sectional count pairs

A.1 REAR END VS. FIXED OBJECT

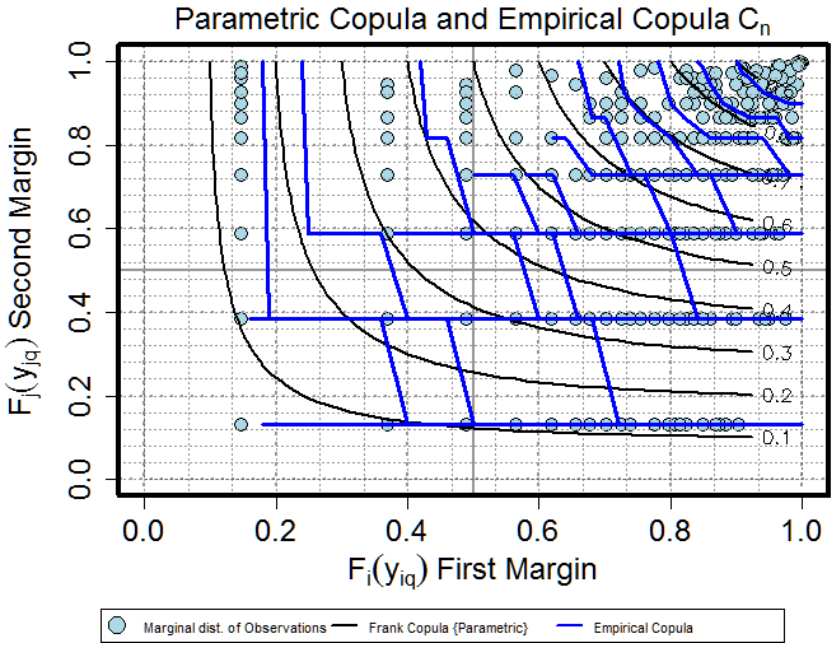


Figure (A.1-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=Rear end vs y=fixed object)

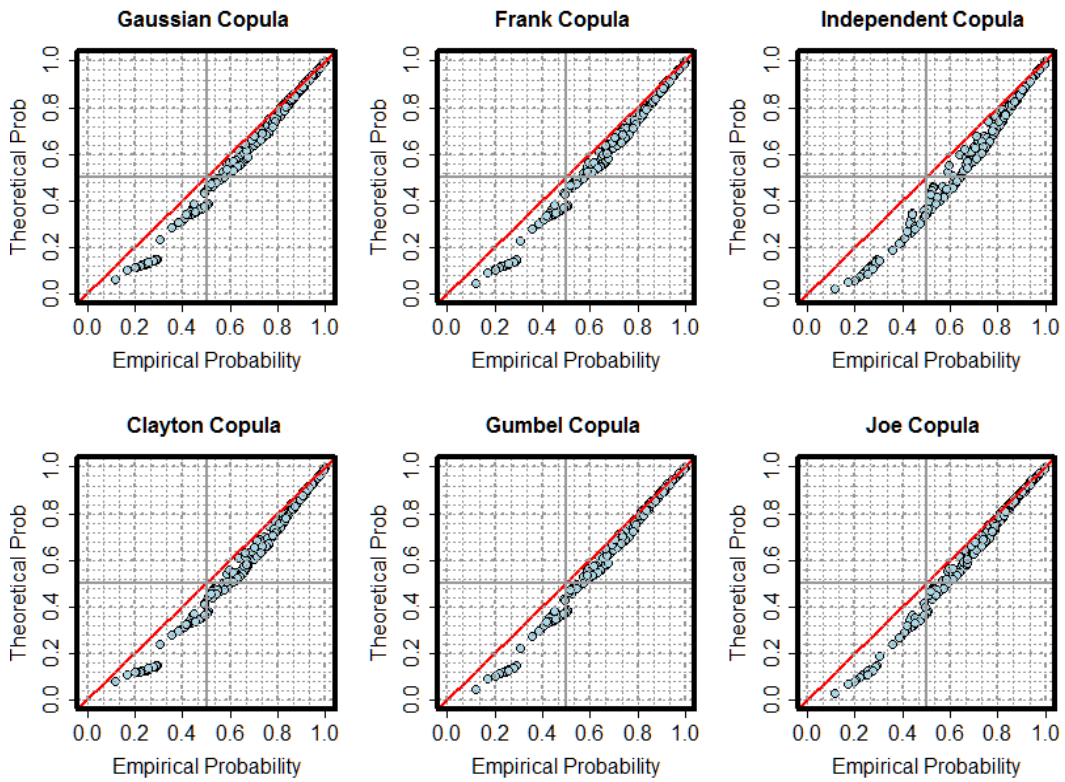


Figure (A.1-2) PP-plot of the parametric copula vs. the empirical copula.

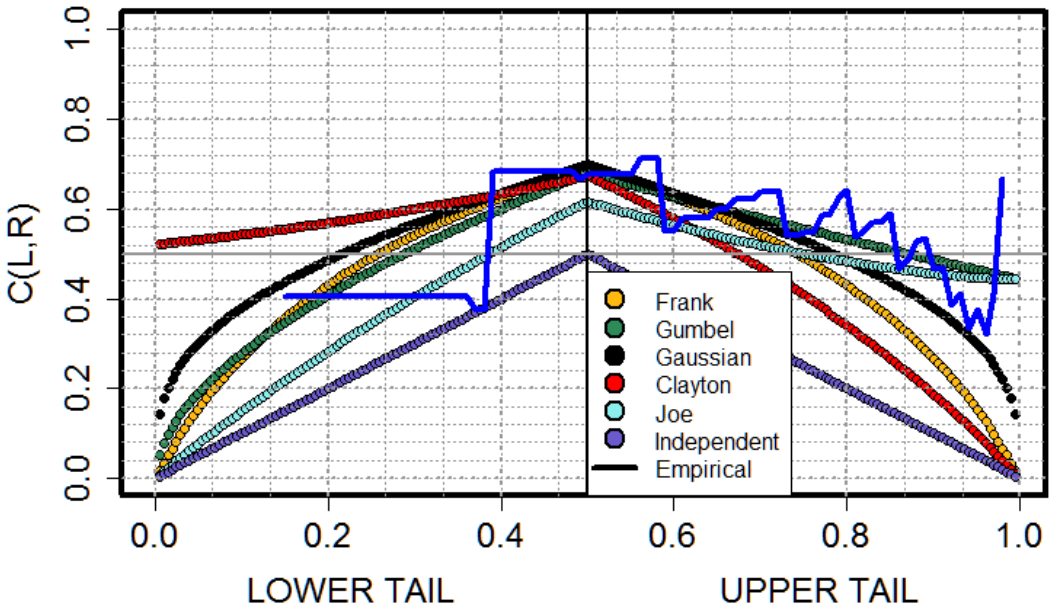


Figure (A.1-3) Tail dependence plot.

A.2 REAR END VS. 'ALL-OTHER'

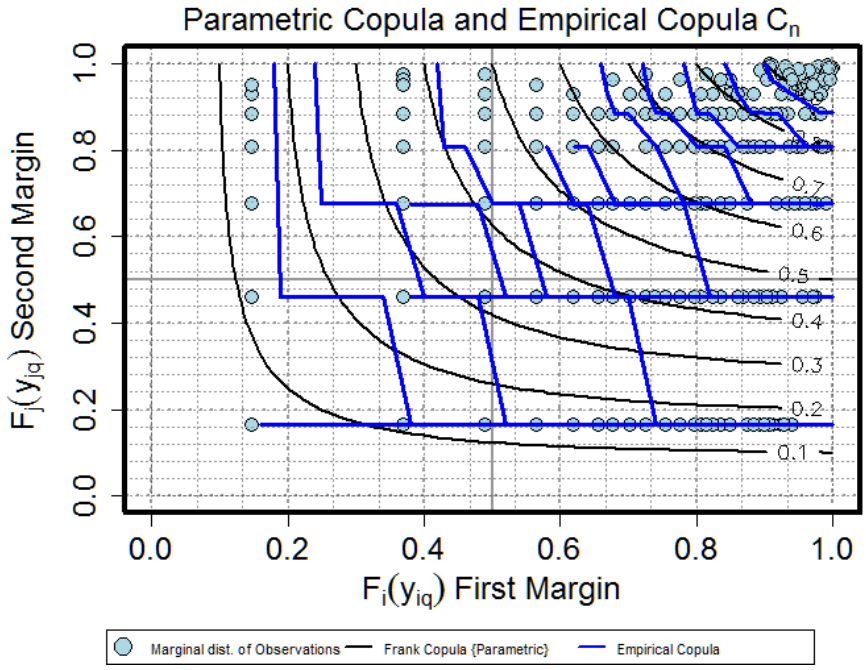


Figure (A.2-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=Rear end vs y='all-other')

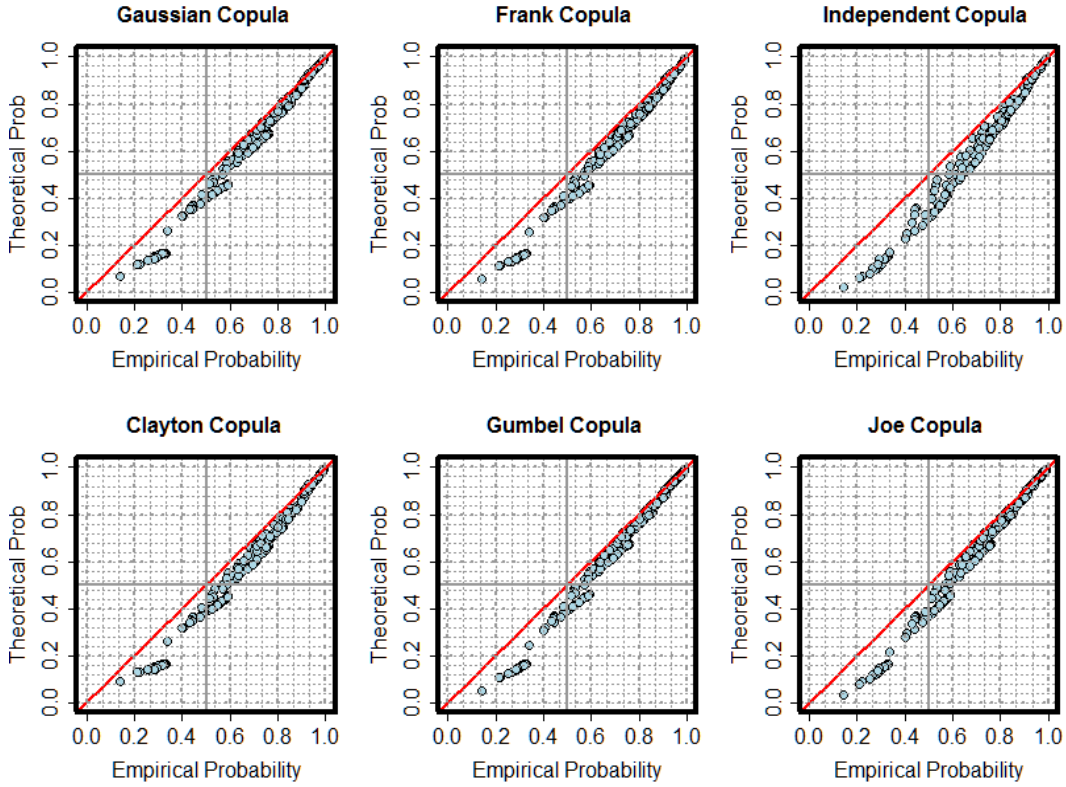


Figure (A.2-2). PP-plot of the parametric copula vs. the empirical copula.

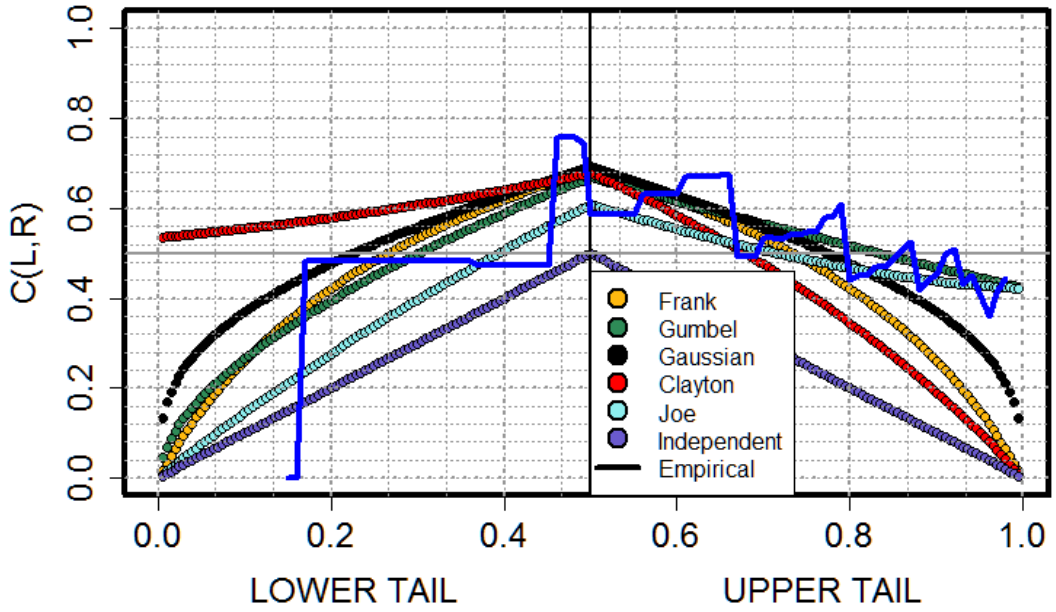


Figure (A.2-3). Tail dependence plot

A.3 SIDESWIPE VS. FIXED OBJECT

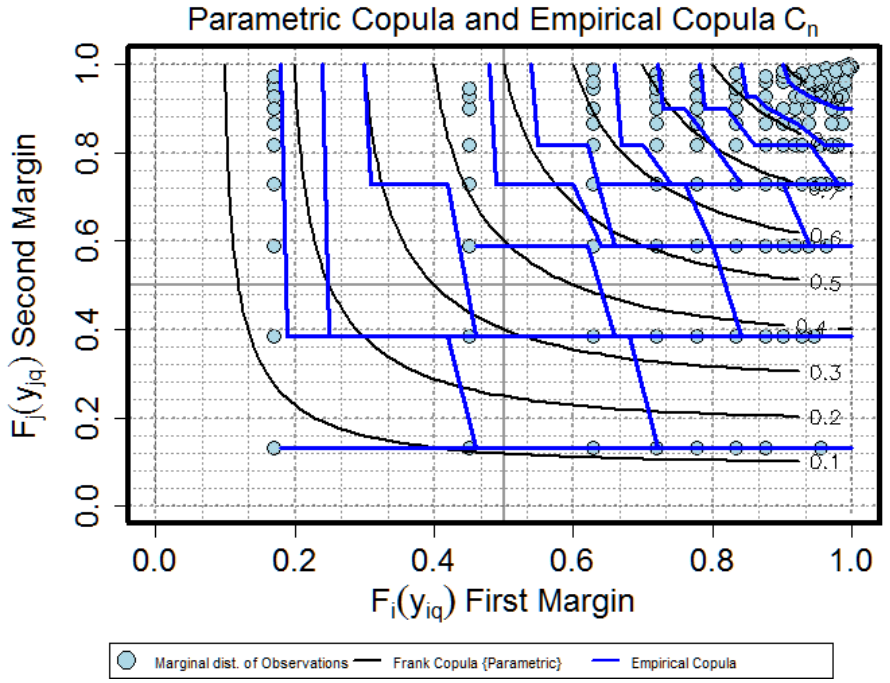


Figure (A.3-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=sideswipe vs y=Fixed object)

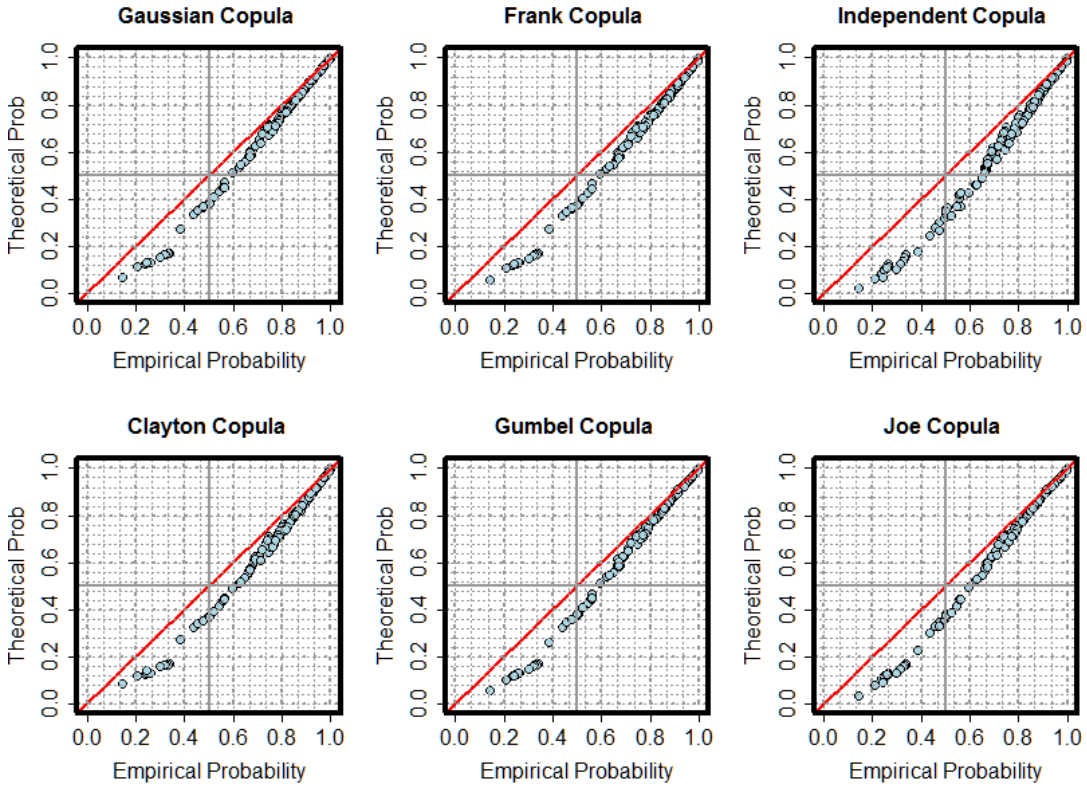


Figure (A.3-2). PP-plot of the parametric copula vs. the empirical copula.

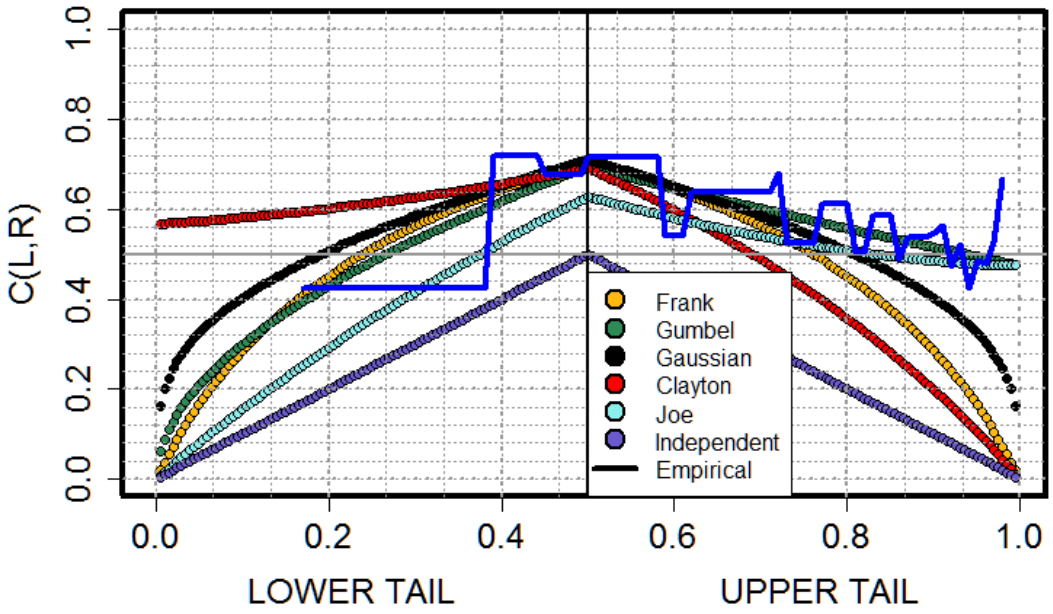


Figure (A.3-3). Tail dependence plot.

A.4 SIDESWIPE VS. 'ALL-OTHER'

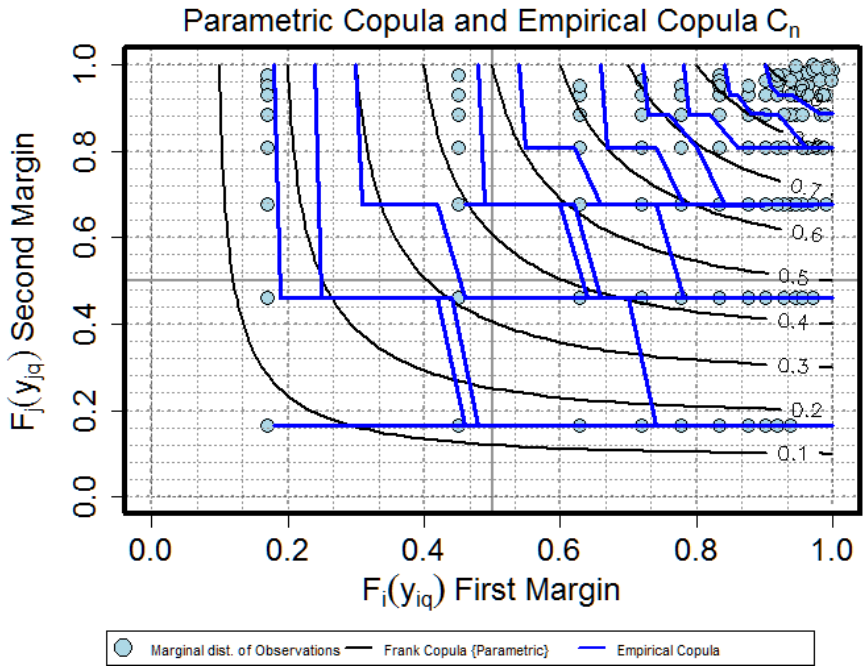


Figure (A.4-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=sideswipe vs y='all-other')

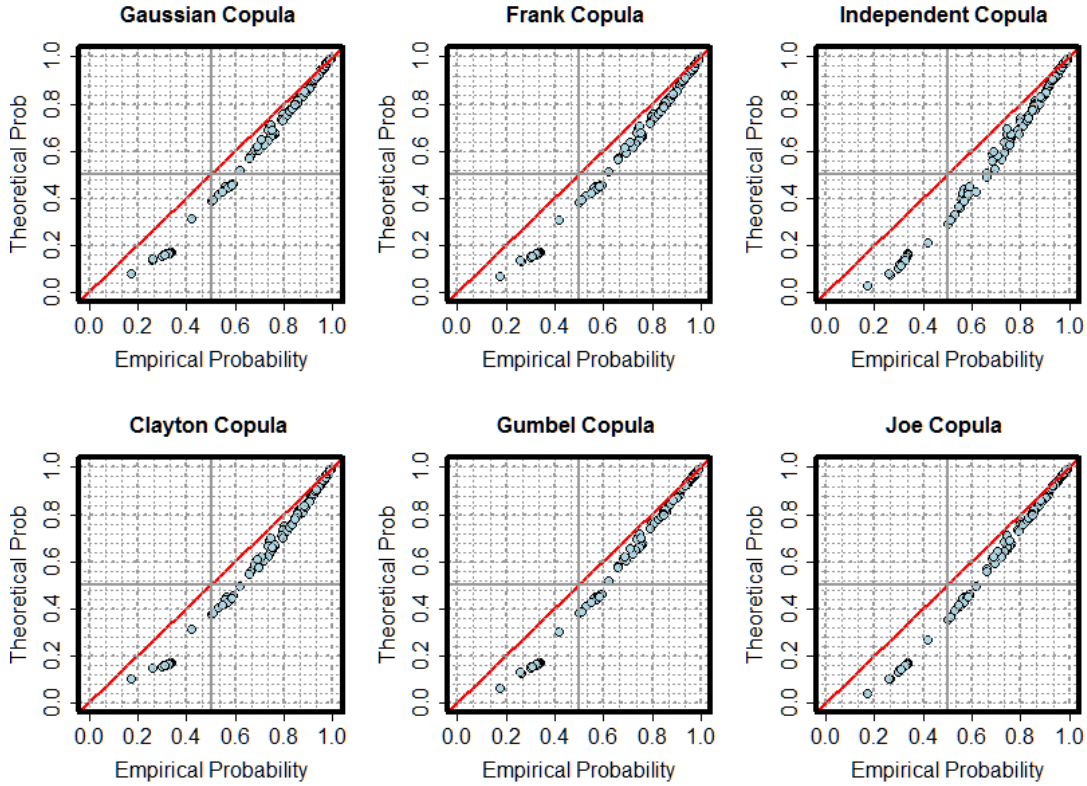


Figure (A.4-2). PP-plot of the parametric copula vs. the empirical copula.

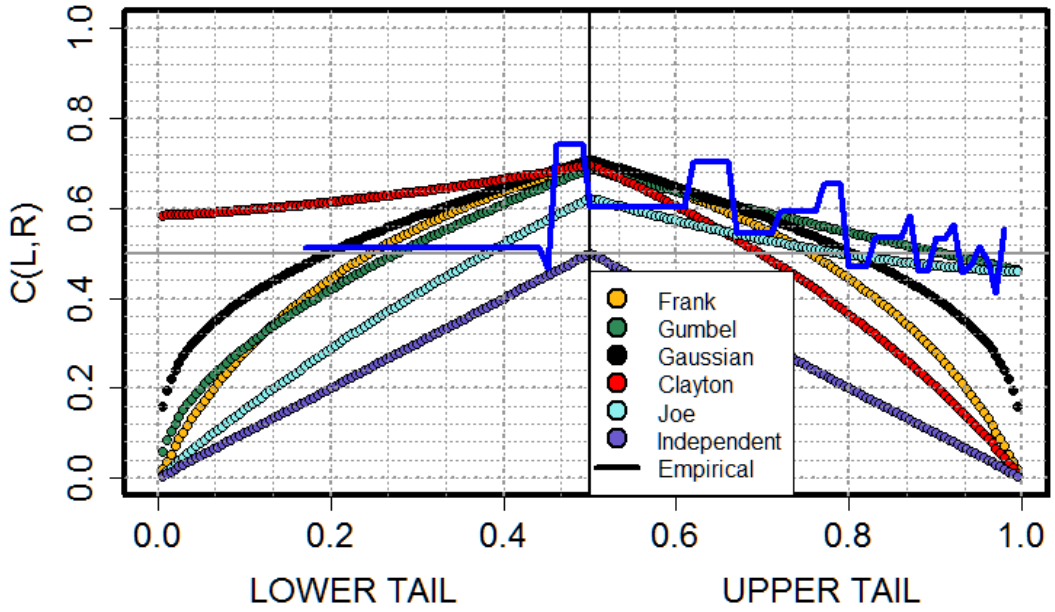


Figure (A.4-3). Tail dependence plot.

A.5 FIXED OBJECT VS. 'ALL-OTHER'

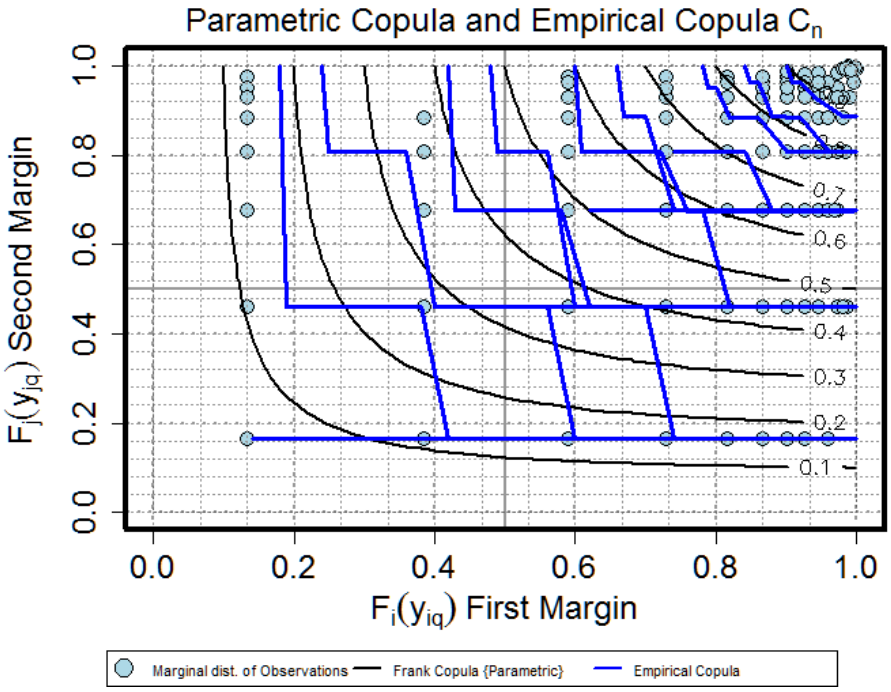


Figure (A.5-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=fixed object vs y='all-other')

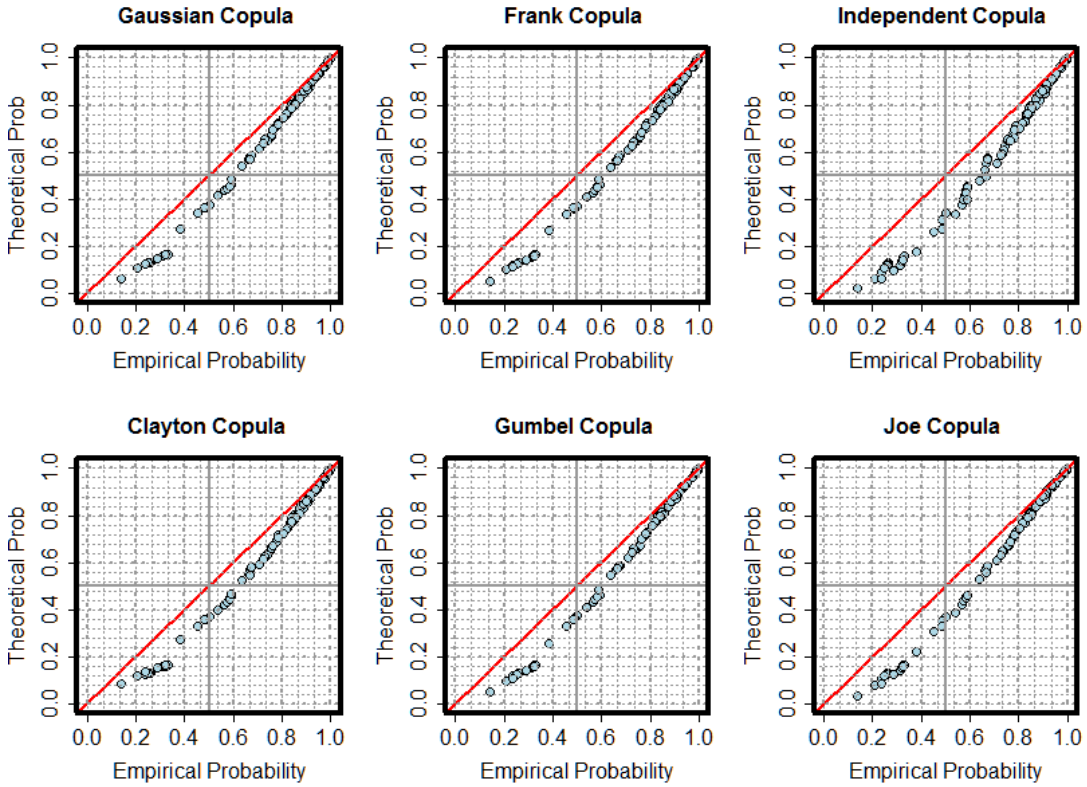


Figure (A.5-2). PP-plot of the parametric copula vs. the empirical copula.

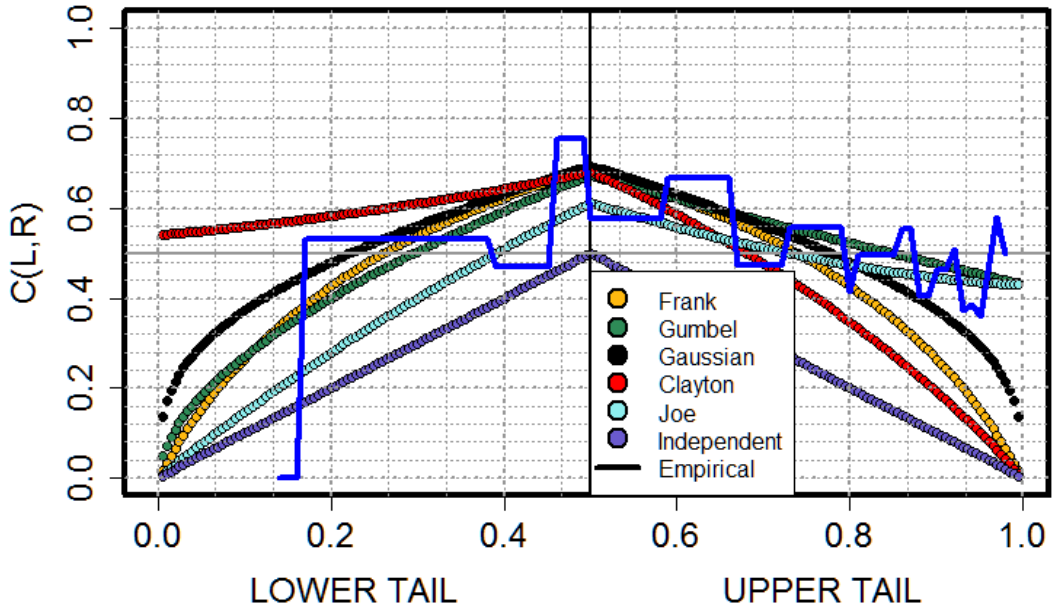


Figure (A.5-3). Tail dependence plot.

Appendix.B

Hoeffding's formula

B.1 DEFINITION

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,

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

For any event π of y_{jq}, one has

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

similarly,

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

for any event μ of y_{iq}. It follows from the identities Eq. (B.1-2) and Eq. (B.1-3) that

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

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

and

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

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

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

which is identical to Hoeffding's formula given in Eq. (5-18) and it can be used to get the covariance between two count dependent random variables.

Appendix.D

Panel count pairs

D.1 REAR-END

D.1.1 Between the year 2005 and 2007

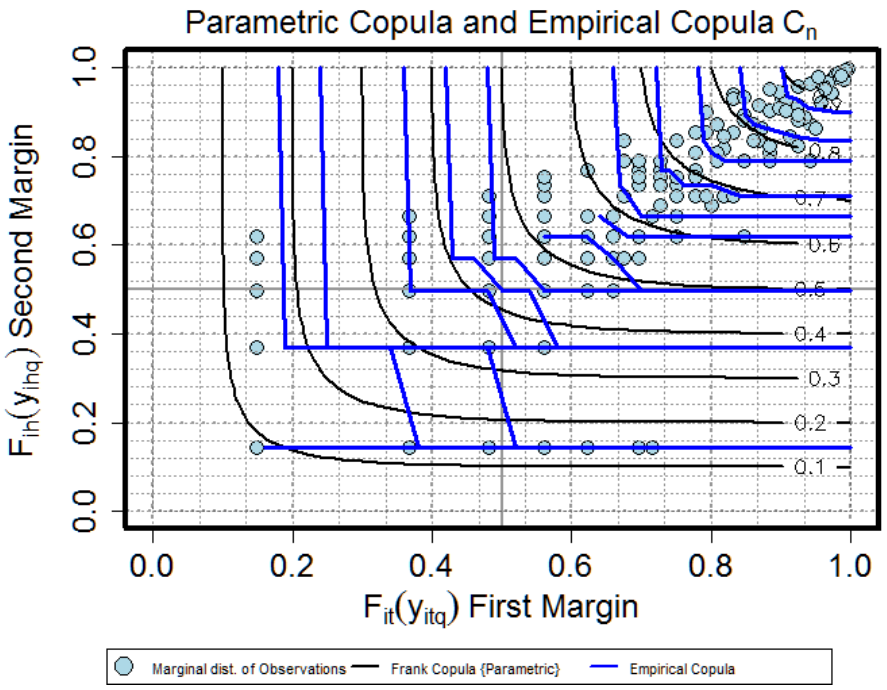


Figure (D.1.1-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=rear-end 2005 vs y=rear-end2007)

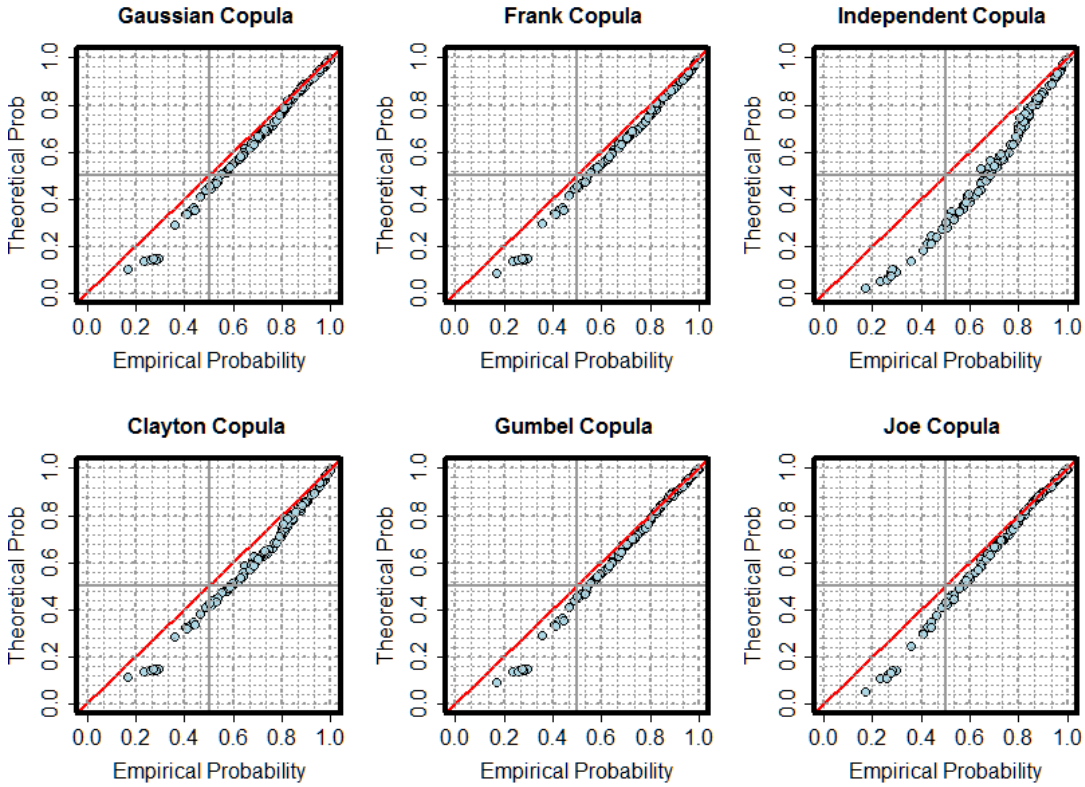


Figure (D.1.1-2). PP-plot of the parametric copula vs. the empirical copula.

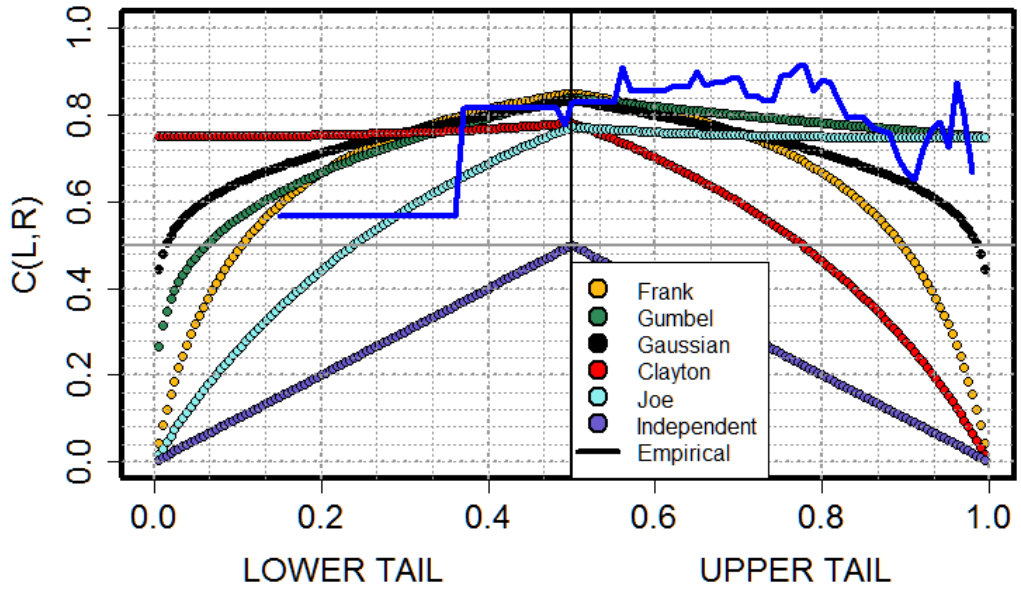


Figure (D.1.1-3). Tail dependence plot.

D.1.2 Between the year 2006 and 2007

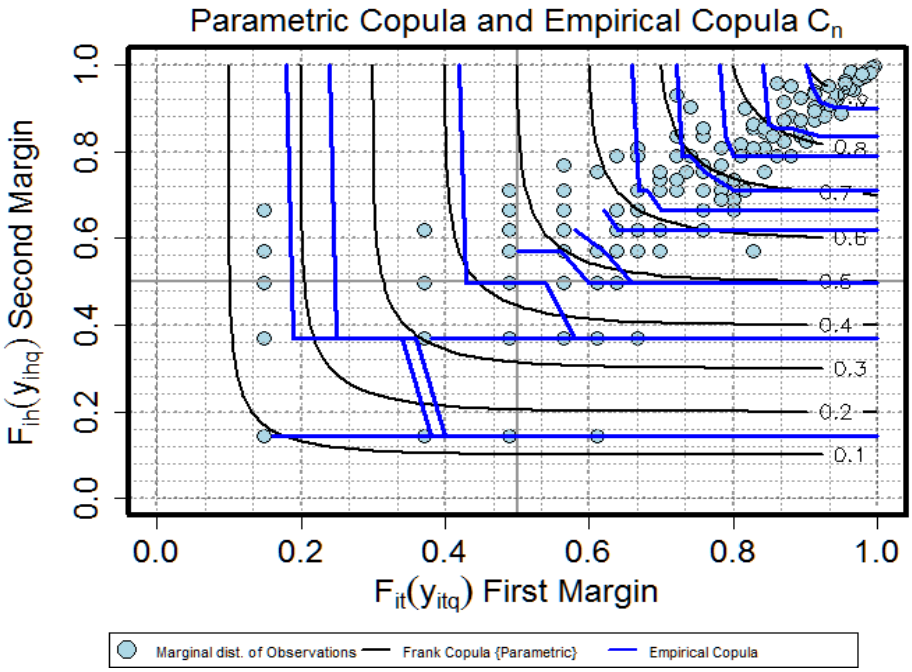


Figure (D.1.2-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=rear-end 2006 vs y=rear-end2007)

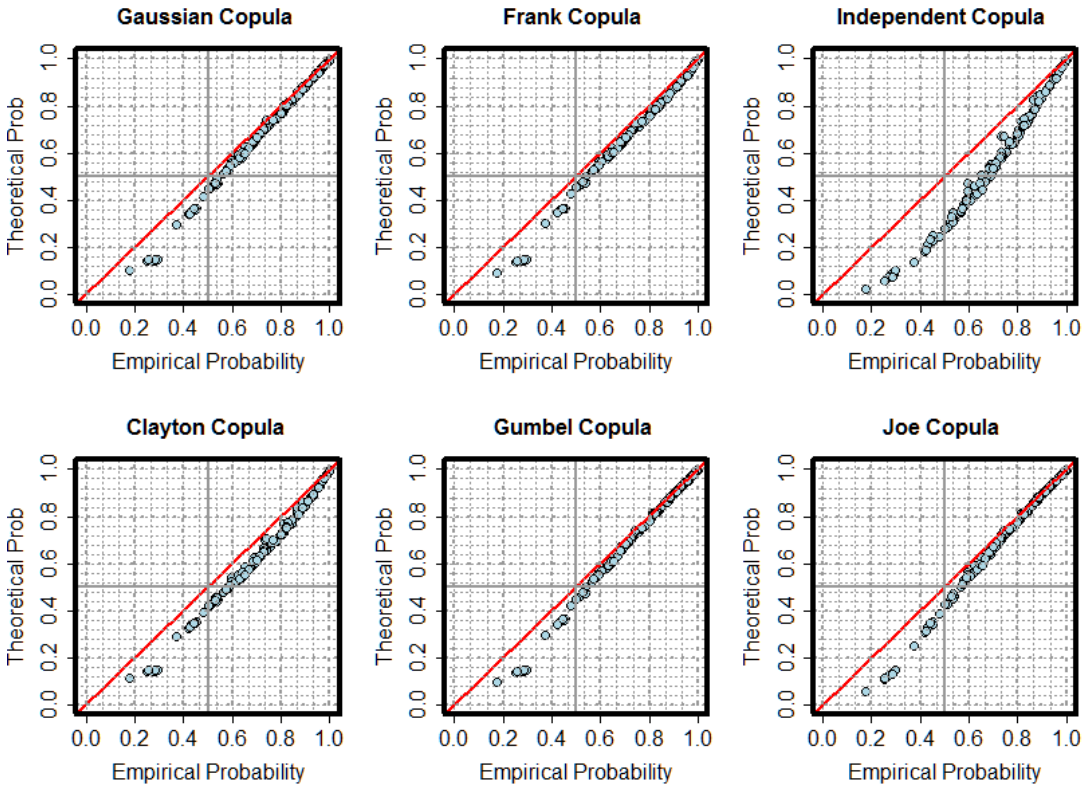


Figure (D.1.2-2). PP-plot of the parametric copula vs. the empirical copula.

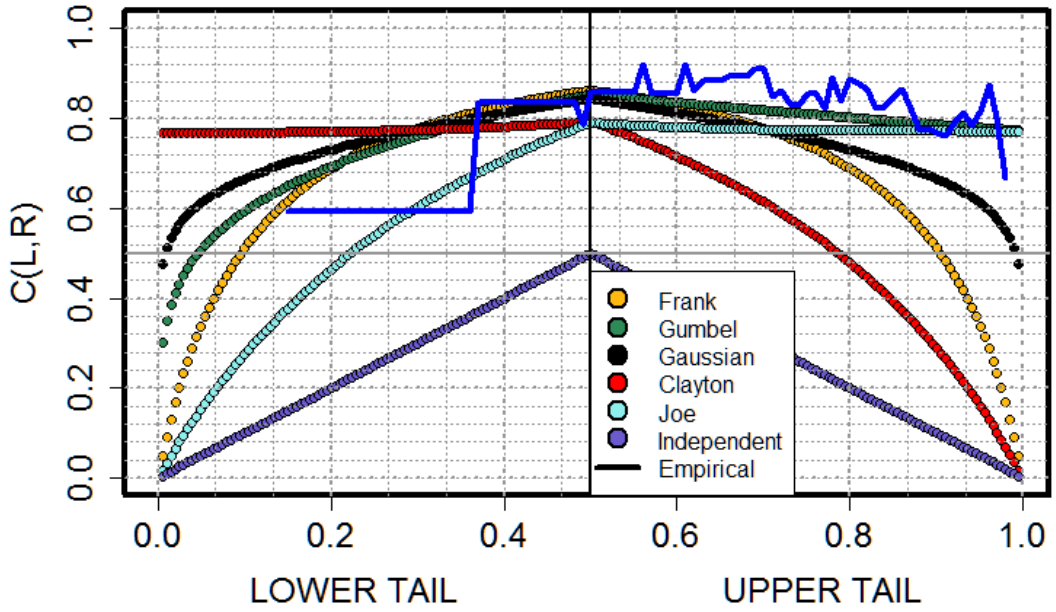


Figure (D.1.2-3). Tail dependence plot.

D.2 SIDESWIPE

D.2.1 Between the year 2005 and 2006

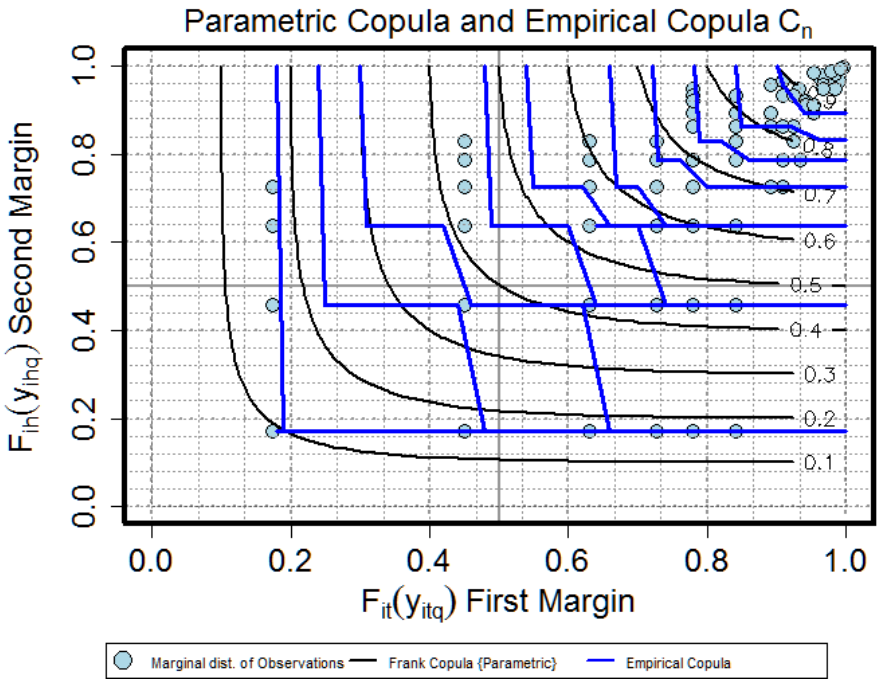


Figure (D.2.1-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=sideswipe 2005 vs y=sideswipe 2006)

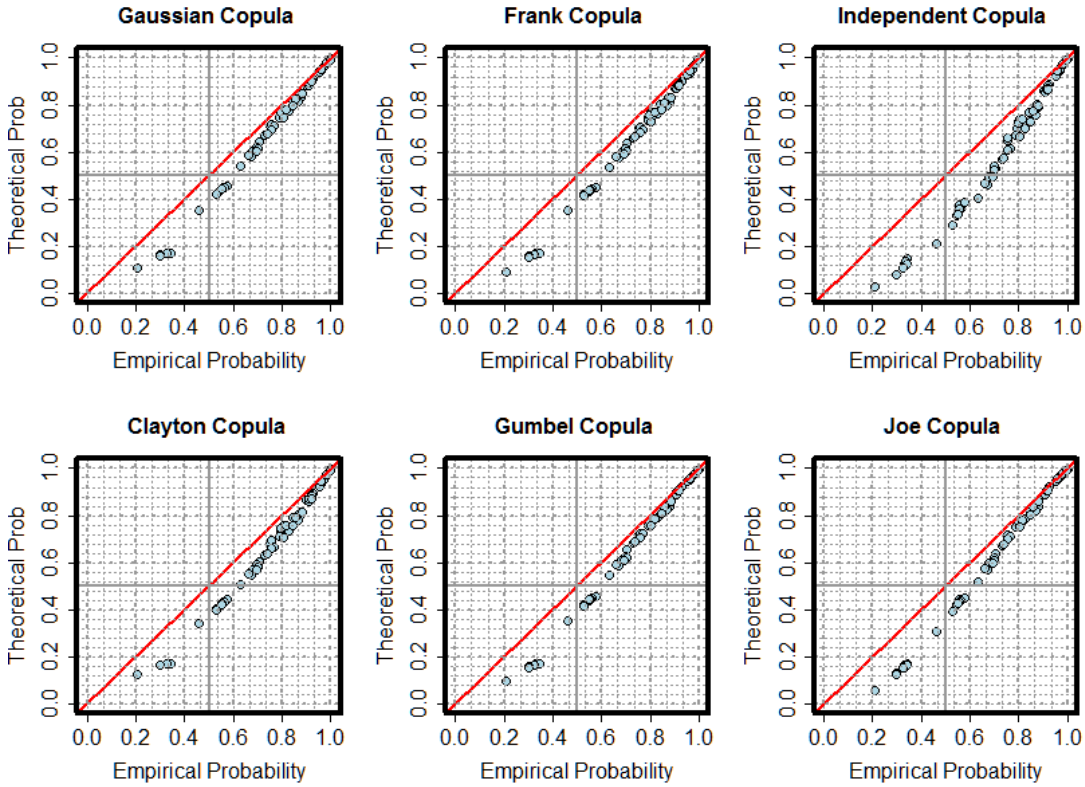


Figure (D.2.1-2). PP-plot of the parametric copula vs. the empirical copula.

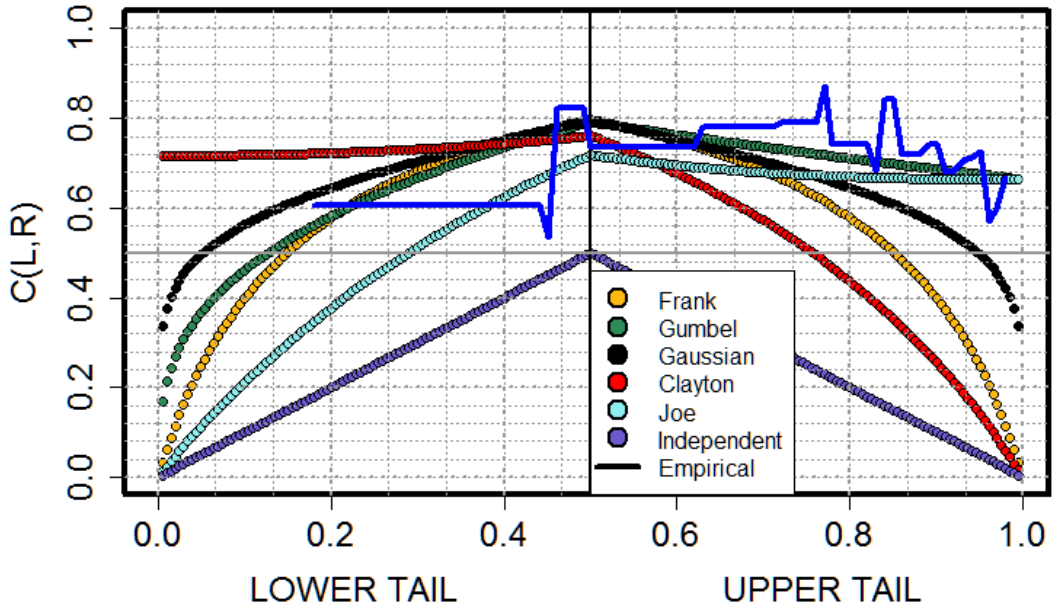


Figure (D.2.1-3). Tail dependence plot.

D.2.2 Between the year 2005 and 2007

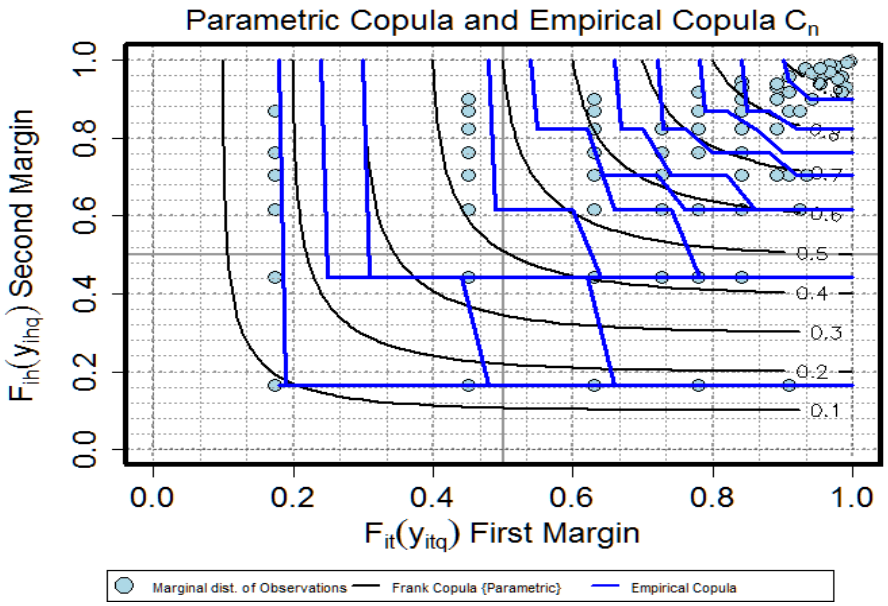


Figure (D.2.2-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=sideswipe 2005 vs y=sideswipe 2007)

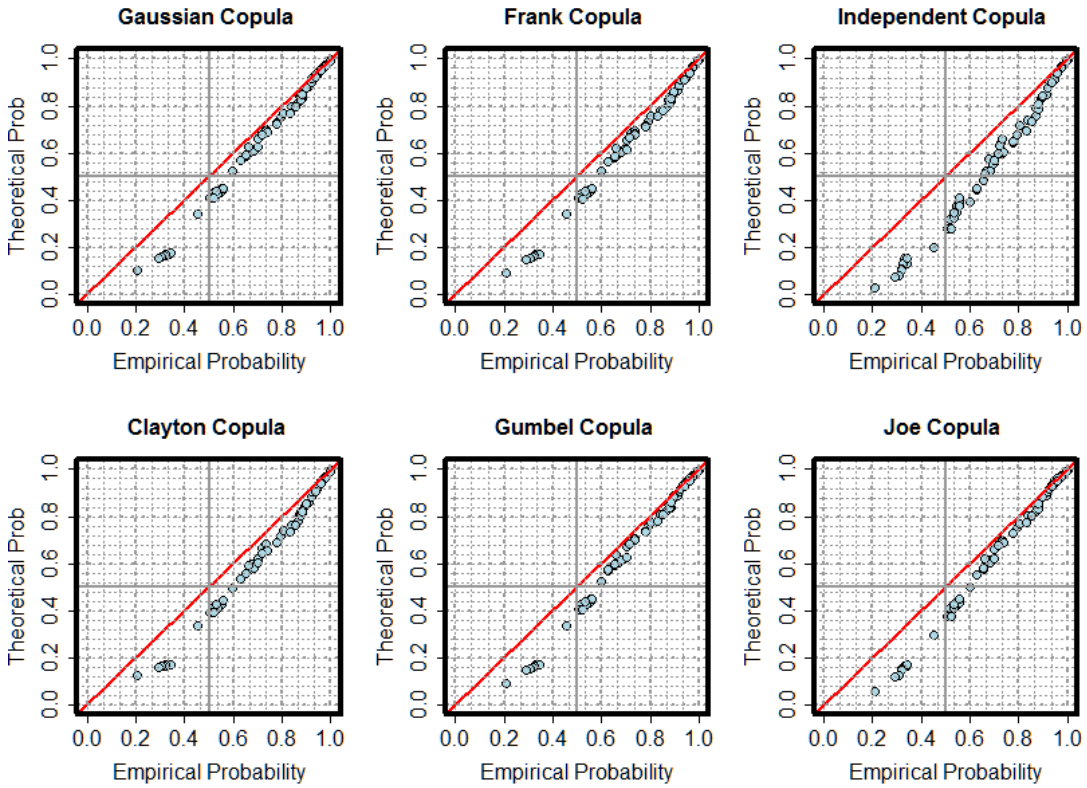


Figure (D.2.2-2). PP-plot of the parametric copula vs. the empirical copula.

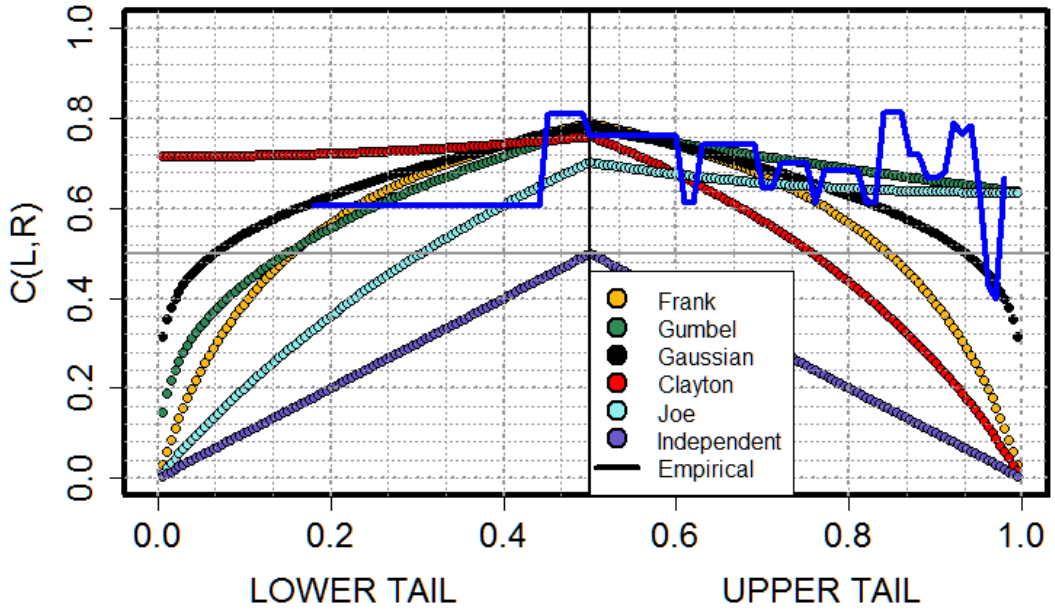


Figure (D.2.2-3). Tail dependence plot.

D.2.3 Between the year 2006 and 2007

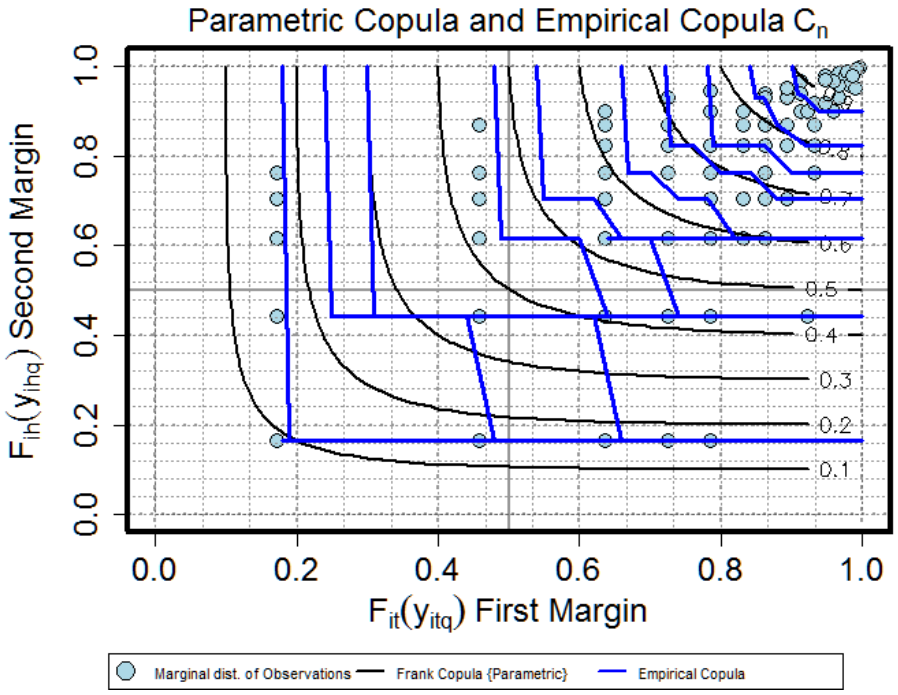


Figure (D.2.3-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=sideswipe 2006 vs y=sideswipe 2007)

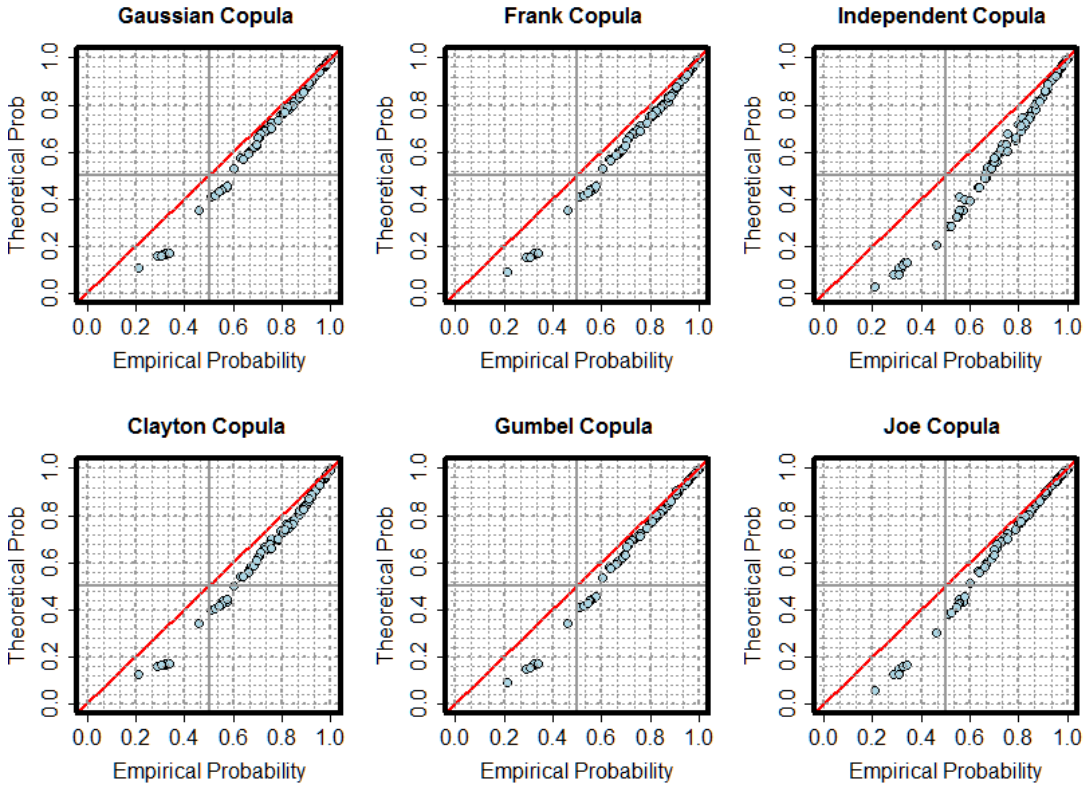


Figure (D.2.3-2). PP-plot of the parametric copula vs. the empirical copula.

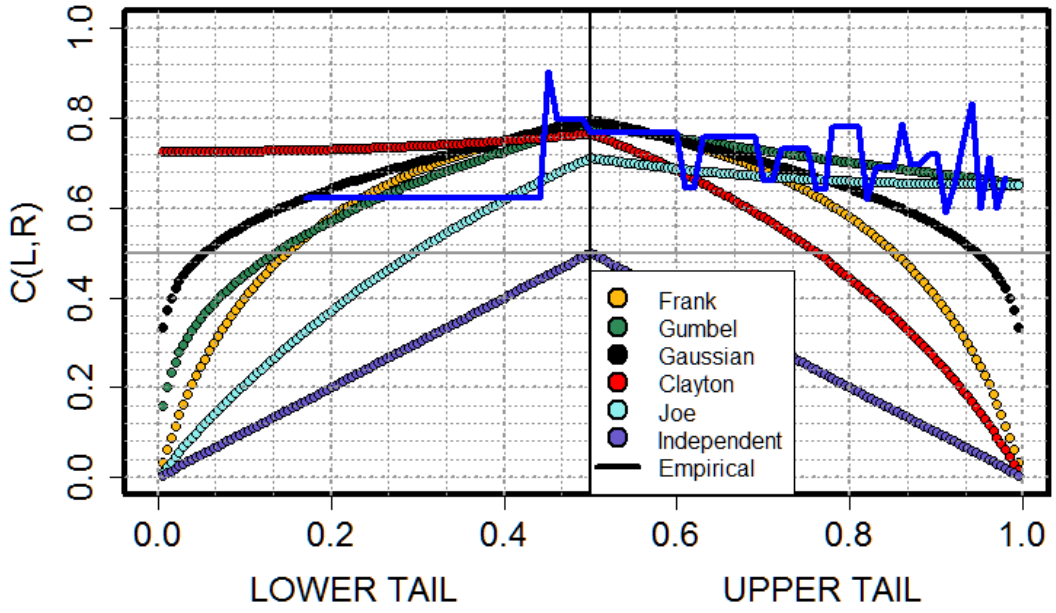


Figure (D.2.3-3). Tail dependence plot.

D.3 FIXED OBJECT

D.3.1 Between the year 2005 and 2006

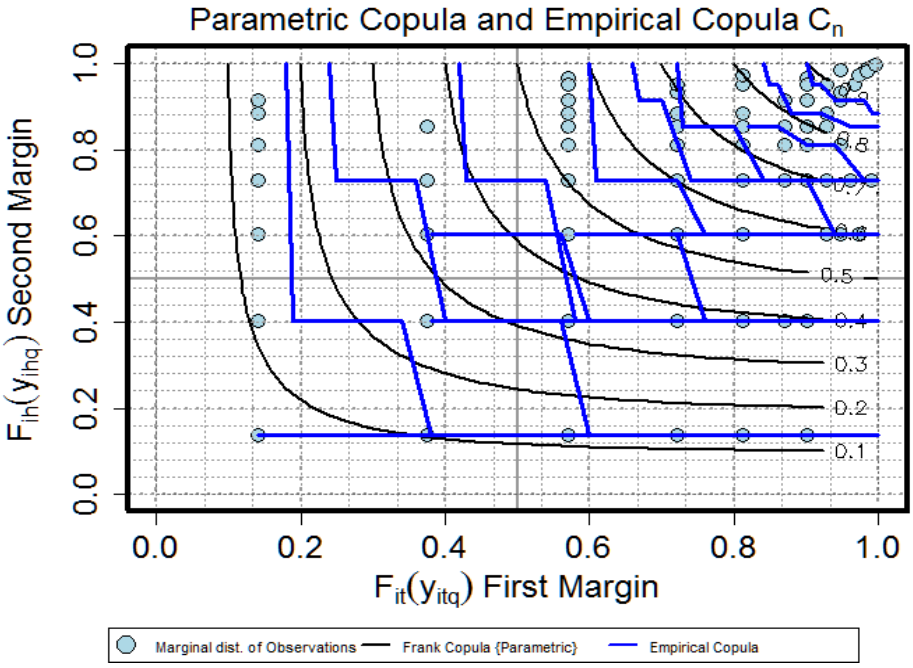


Figure (D.3.1-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=fixed object 2005 vs y=fixed object 2007)

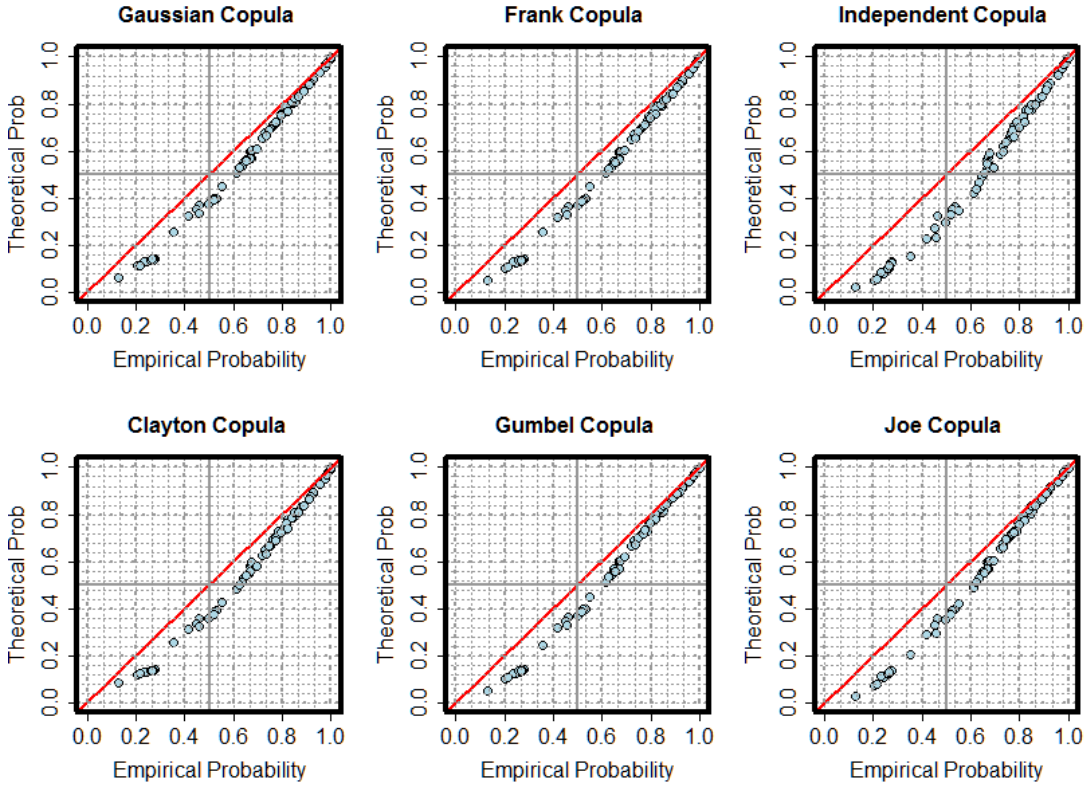


Figure (D.3.1-2). PP-plot of the parametric copula vs. the empirical copula.

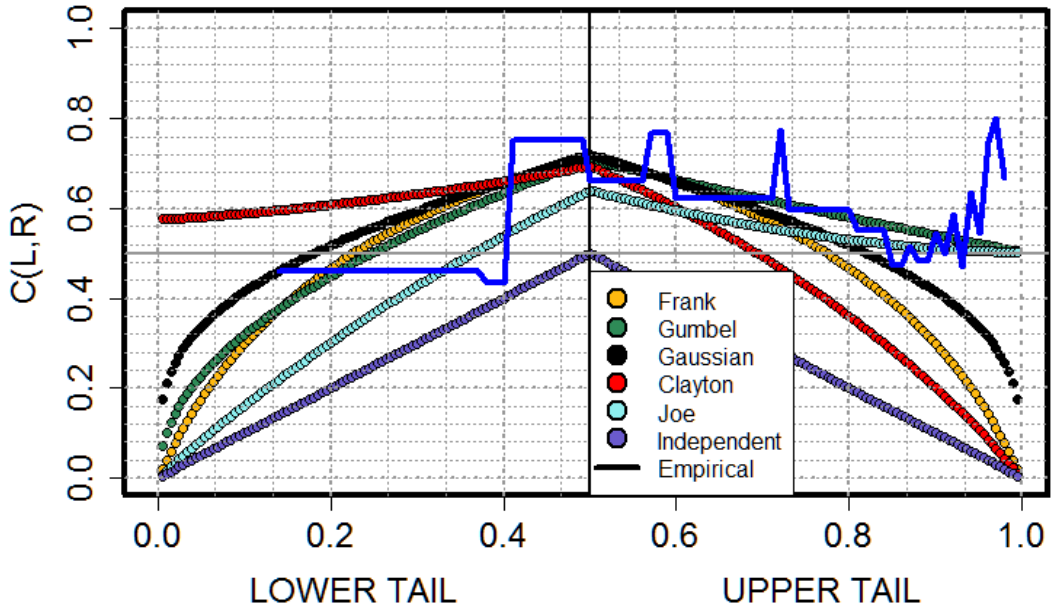


Figure (D.3.1-3). Tail dependence plot.

D.3.2 Between the year 2005 and 2007

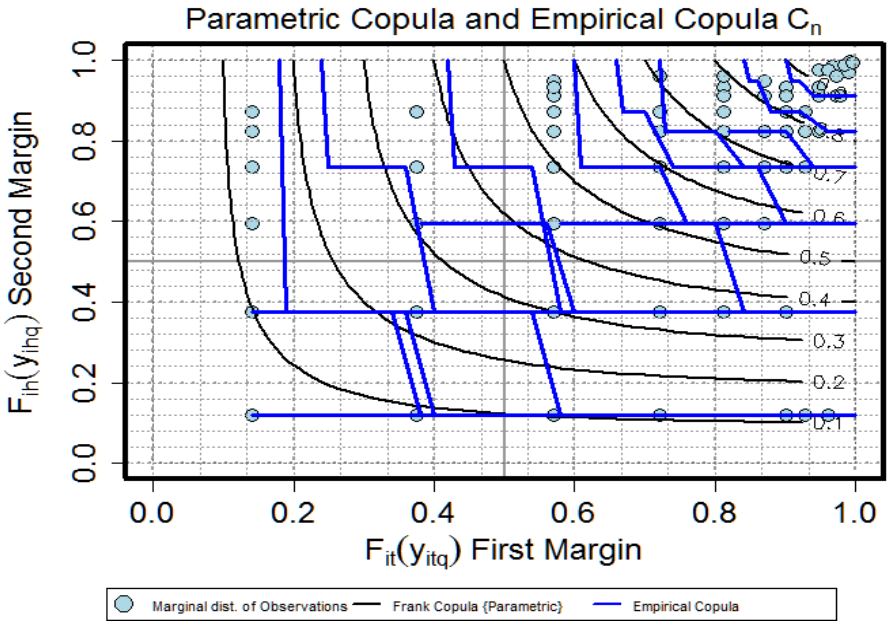


Figure (D.3.2-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=fixed object 2005 vs y=fixed object 2007)

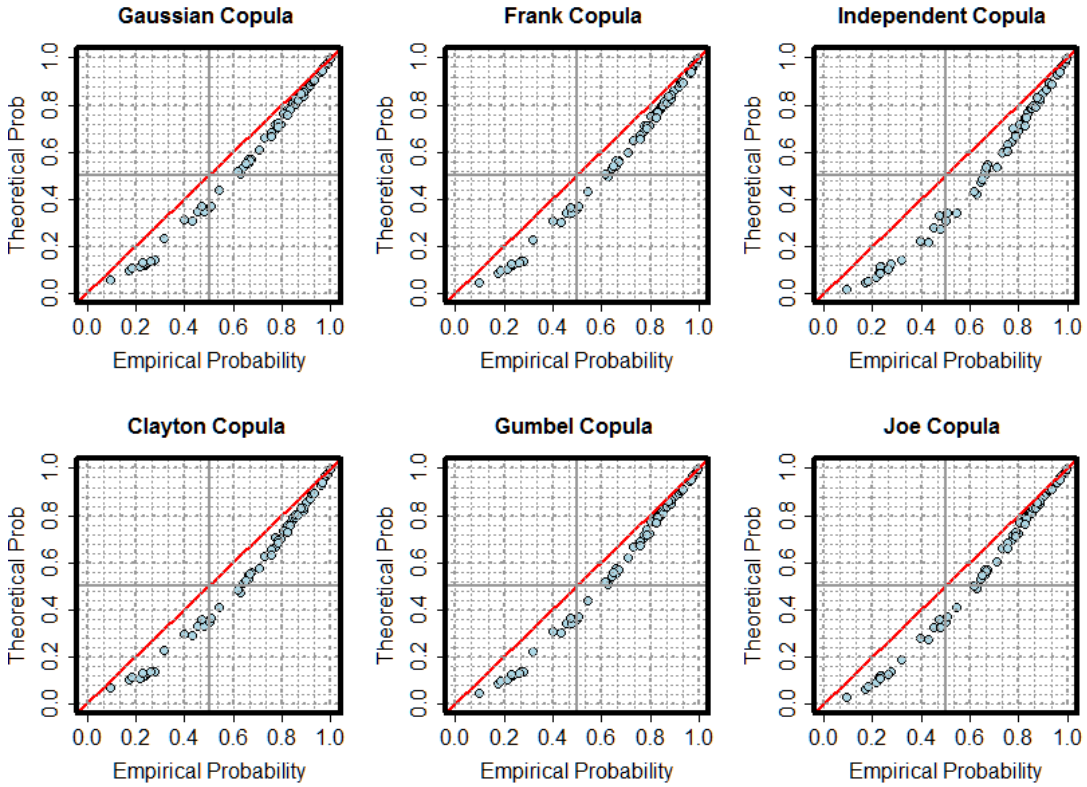


Figure (D.3.2-2). PP-plot of the parametric copula vs. the empirical copula.

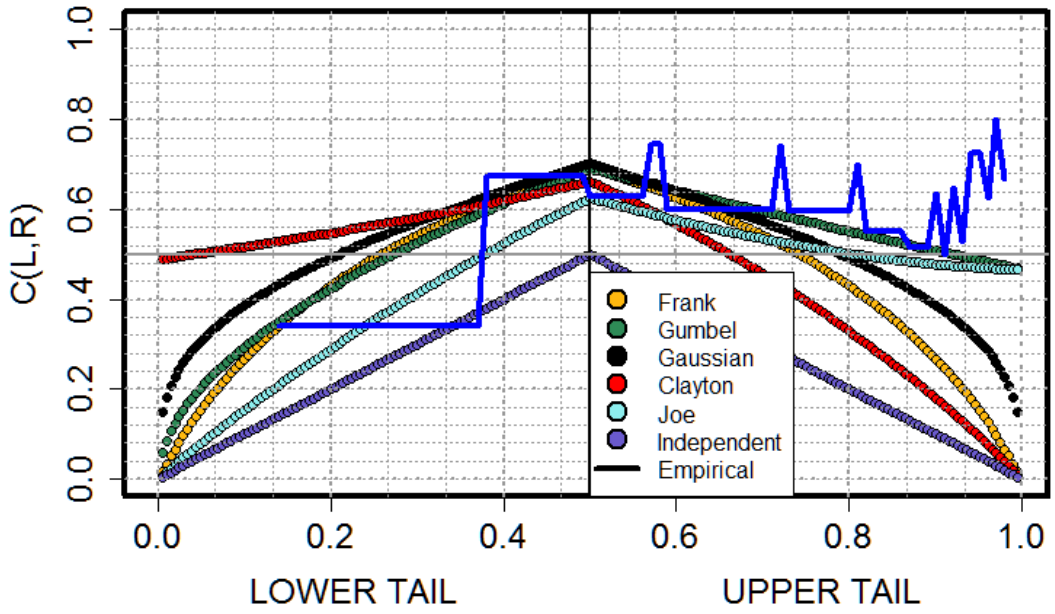


Figure (D.3.2-3). Tail dependence plot.

D.3.3 Between the year 2006 and 2007

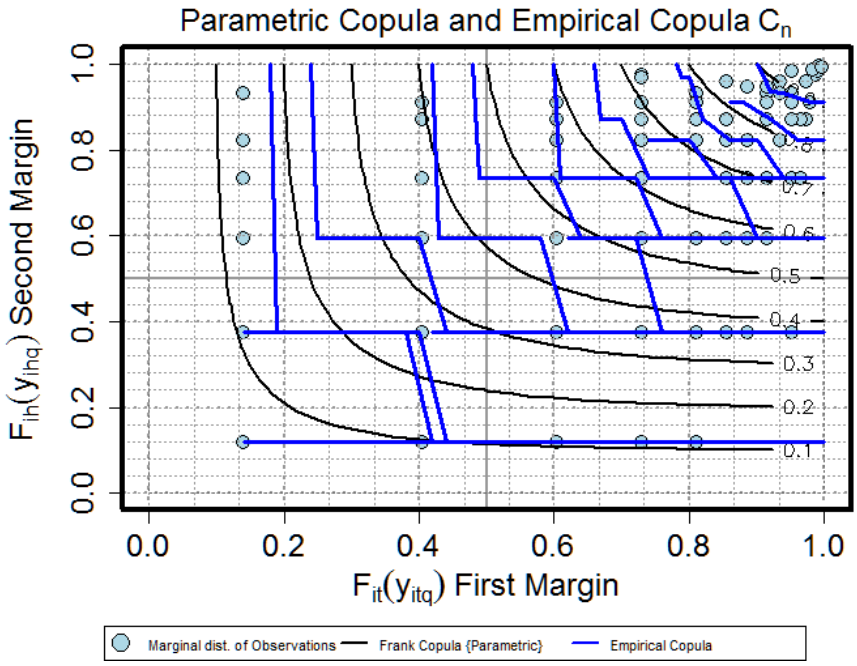


Figure (D.3.3-1). The empirical copula using 1/Q type compared to a selected parametric copula (x=fixed object 2006 vs y=fixed object 2007)

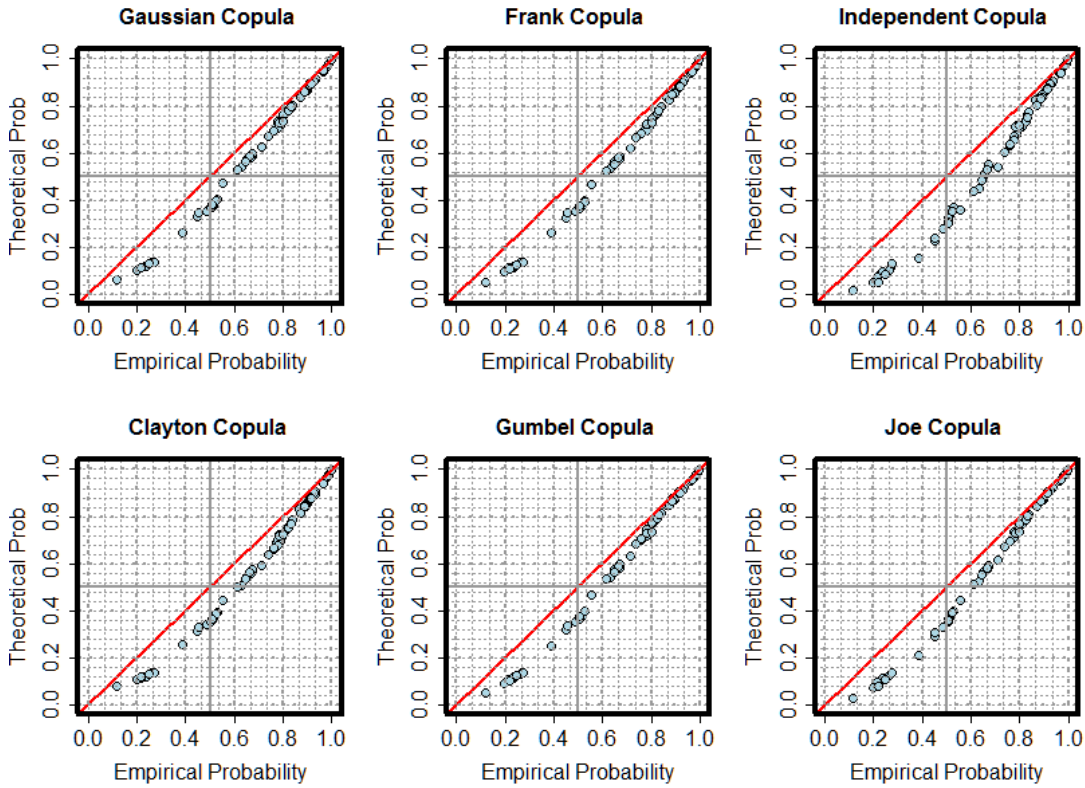


Figure (D.3.3-2). PP-plot of the parametric copula vs. the empirical copula.

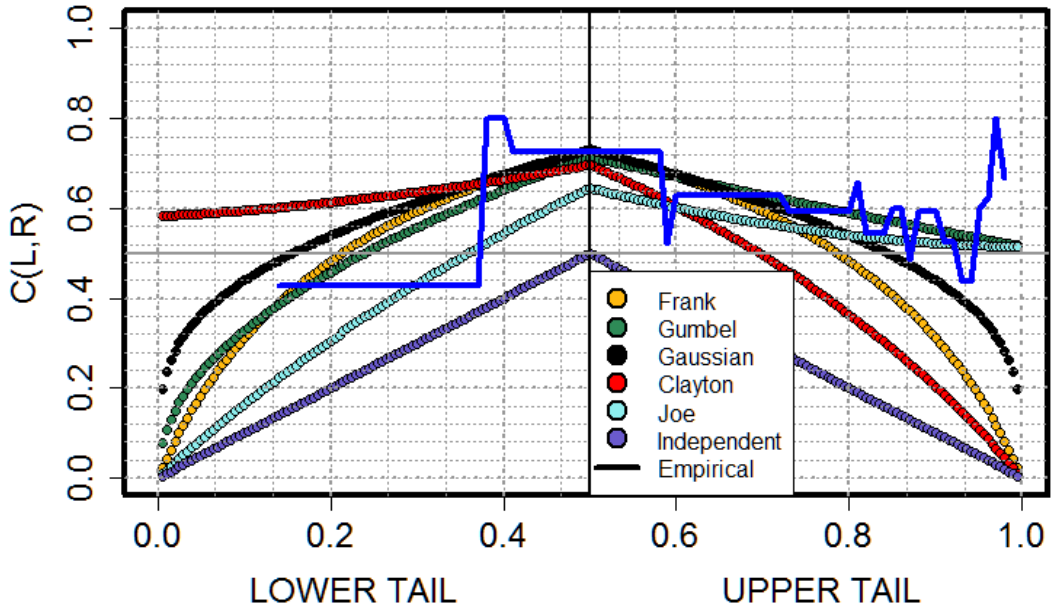


Figure (D.3.3-3). Tail dependence plot.

D.4 'ALL-OTHER'

D.4.1 Between the year 2005 and 2006

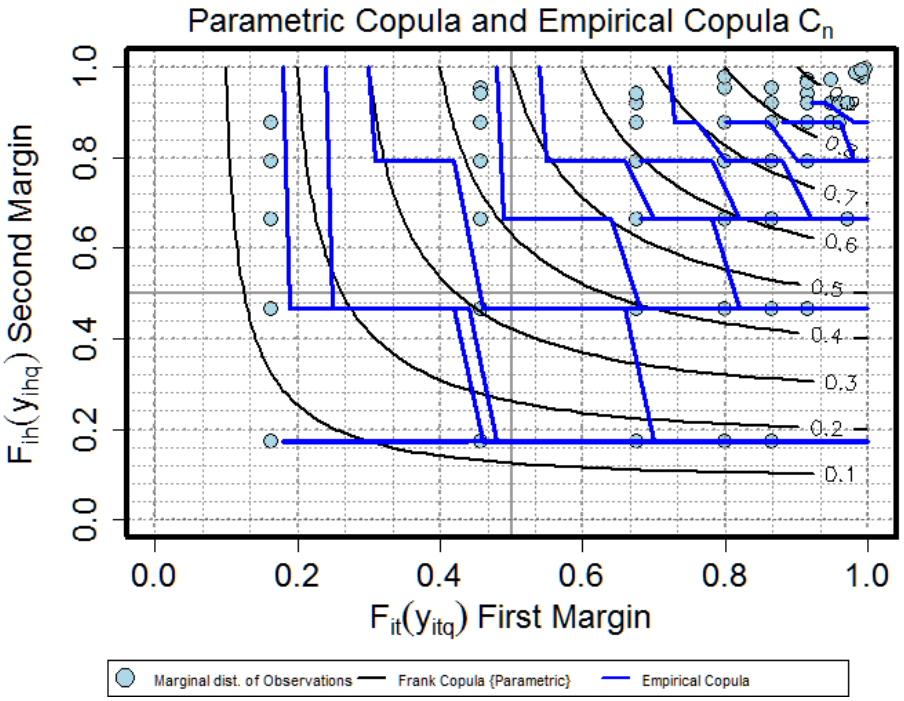


Figure (D.4.1-1). The empirical copula using 1/Q type compared to a selected parametric copula (x='all-other' 2005 vs y='all-other' 2006)

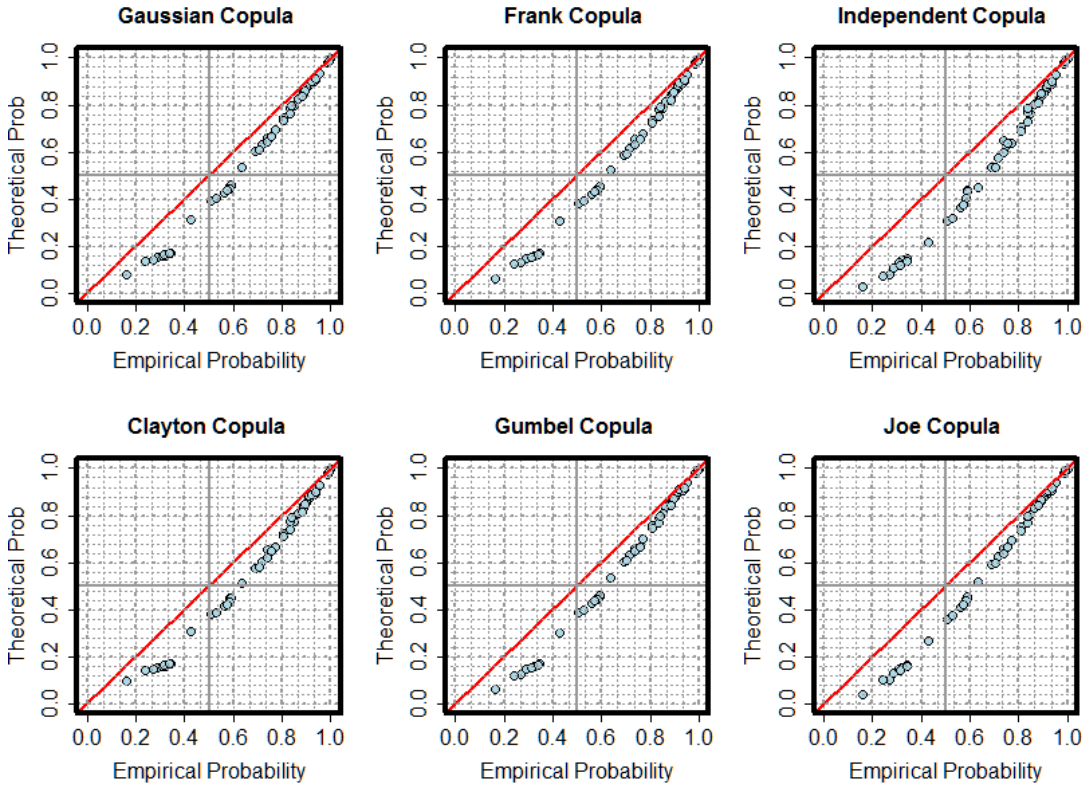


Figure (D.4.1-2). PP-plot of the parametric copula vs. the empirical copula.

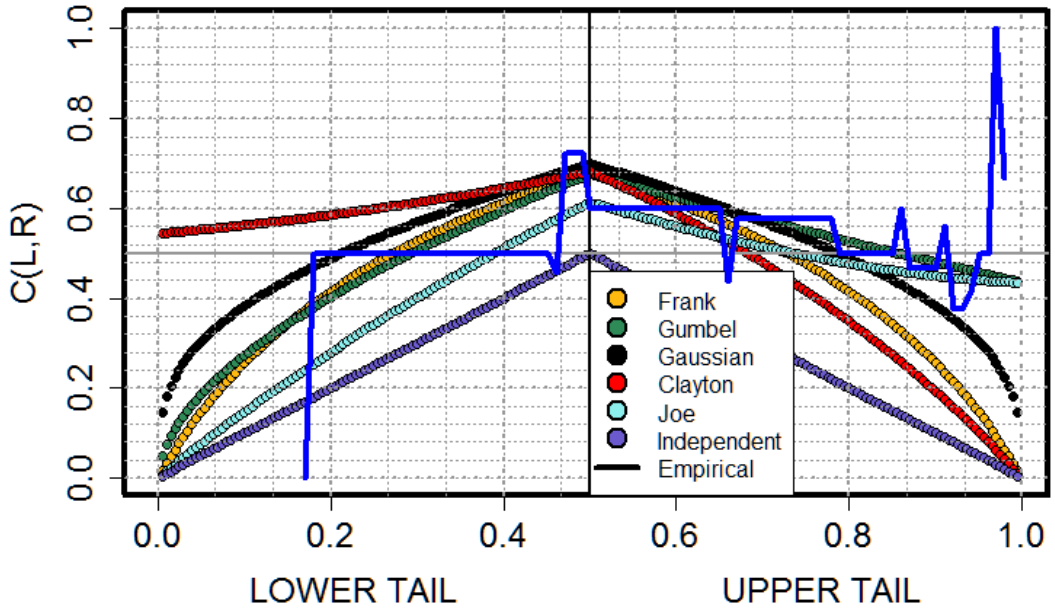


Figure (D.4.1-3). Tail dependence plot.

D.4.2 Between the year 2005 and 2007

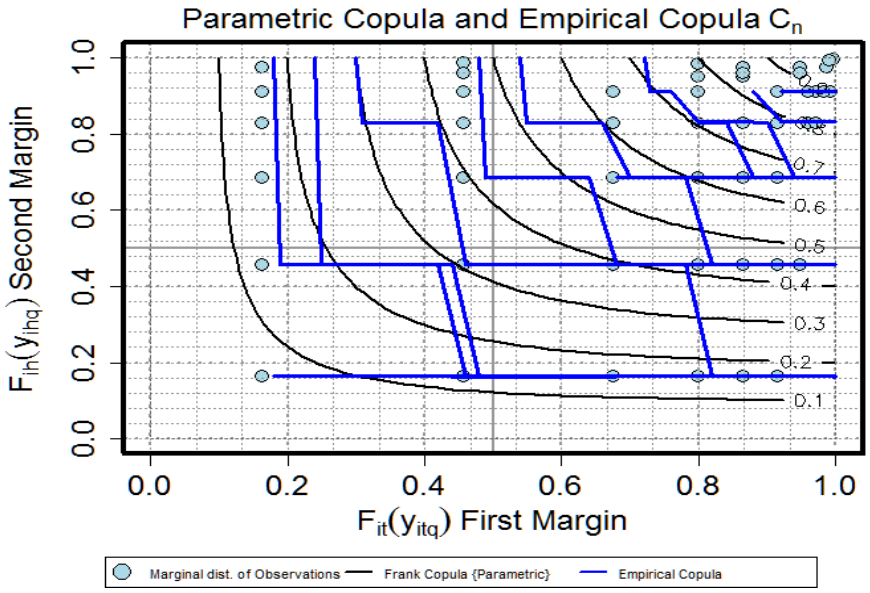


Figure (D.4.2-1). The empirical copula using 1/Q type compared to a selected parametric copula (x='all-other' 2005 vs y='all-other' 2007)

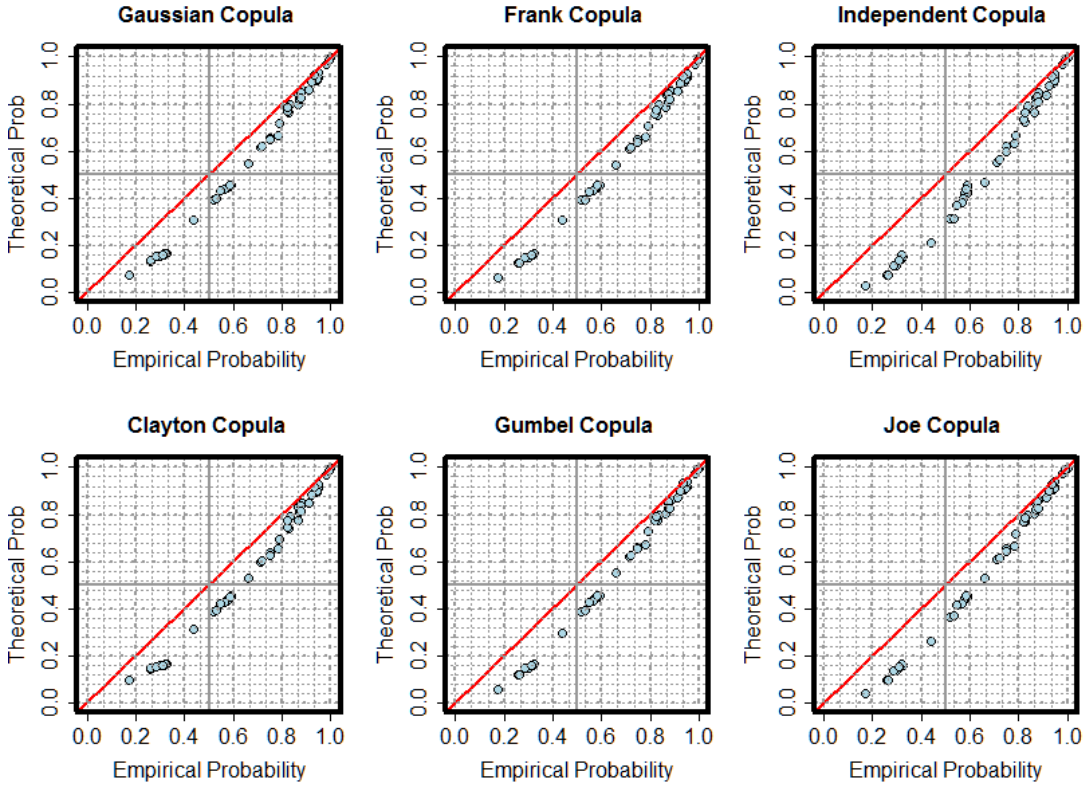


Figure (D.4.2-2). PP-plot of the parametric copula vs. the empirical copula.

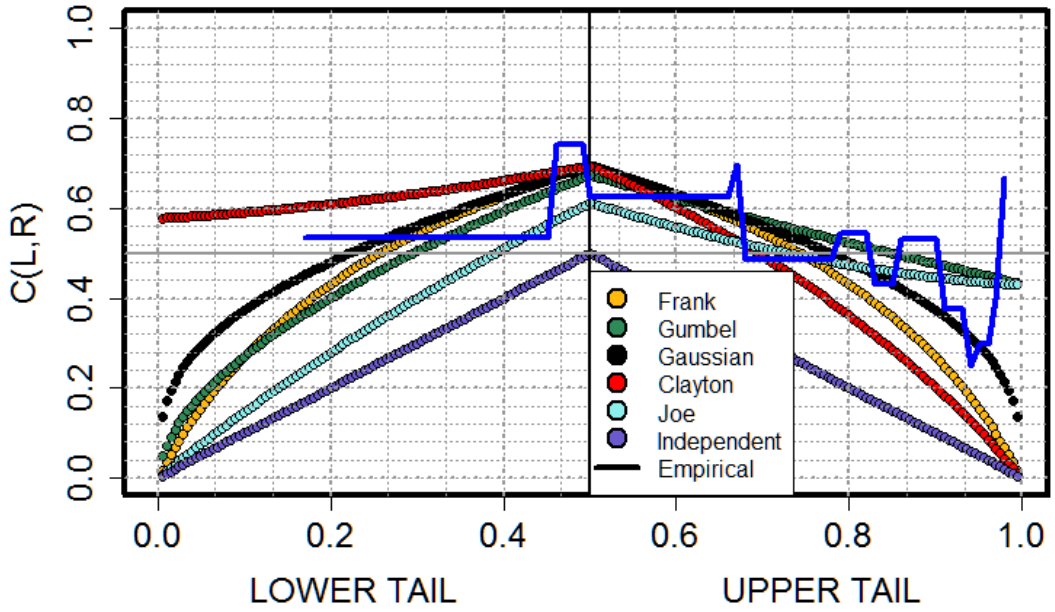


Figure (D.4.2-3). Tail dependence plot.

D.4.3 Between the year 2006 and 2007

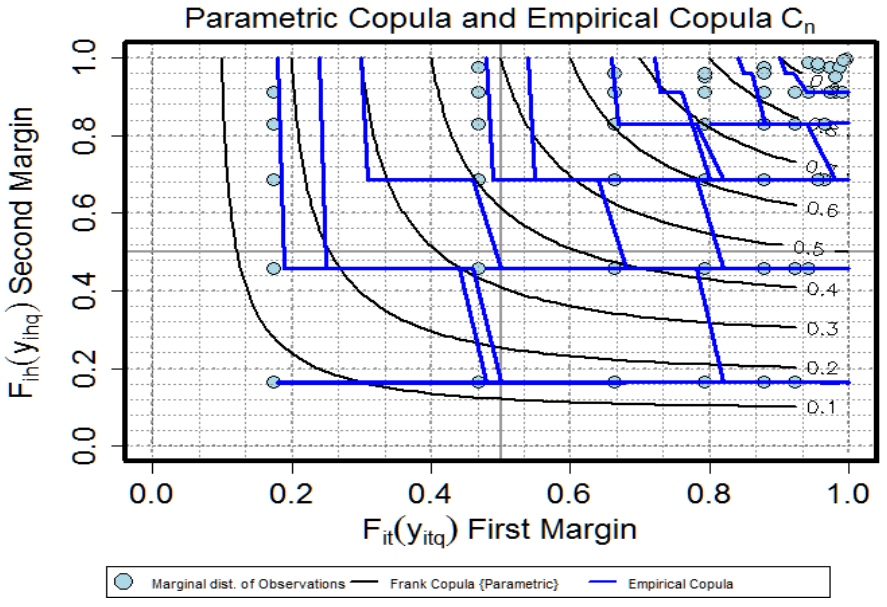


Figure (D.4.3-1). The empirical copula using 1/Q type compared to a selected parametric copula (x='all-other' 2006 vs y='all-other' 2007)

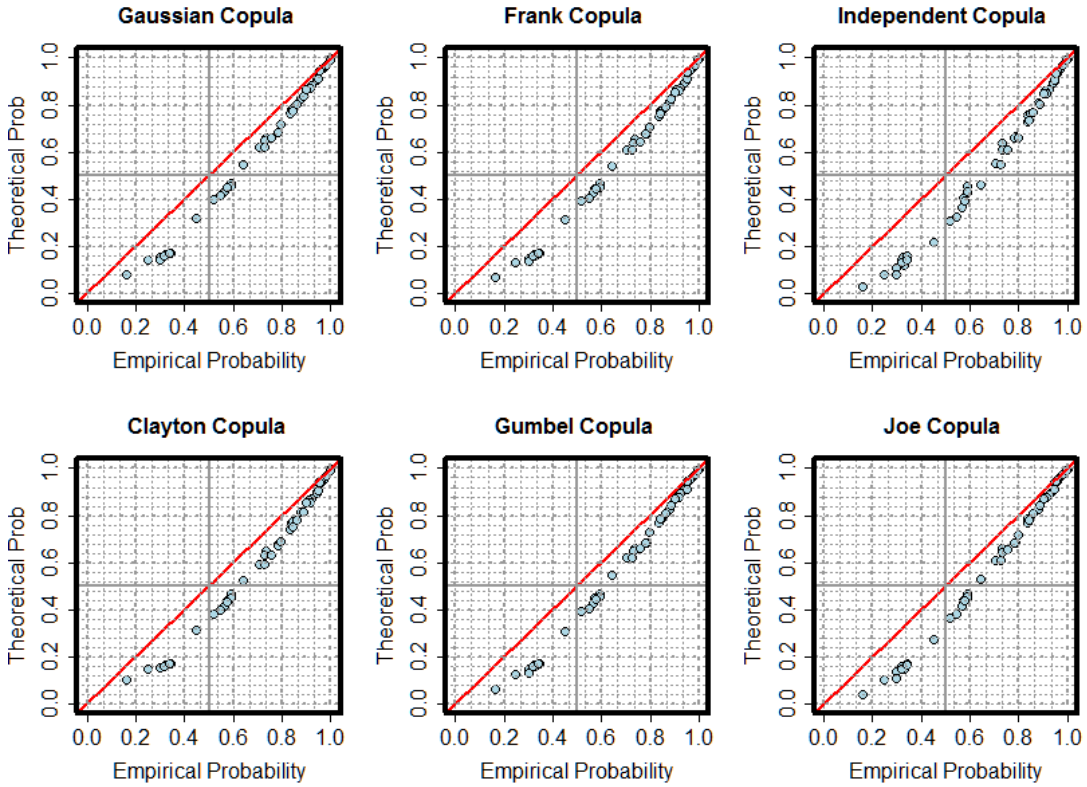


Figure (D.4.3-2). PP-plot of the parametric copula vs. the empirical copula.

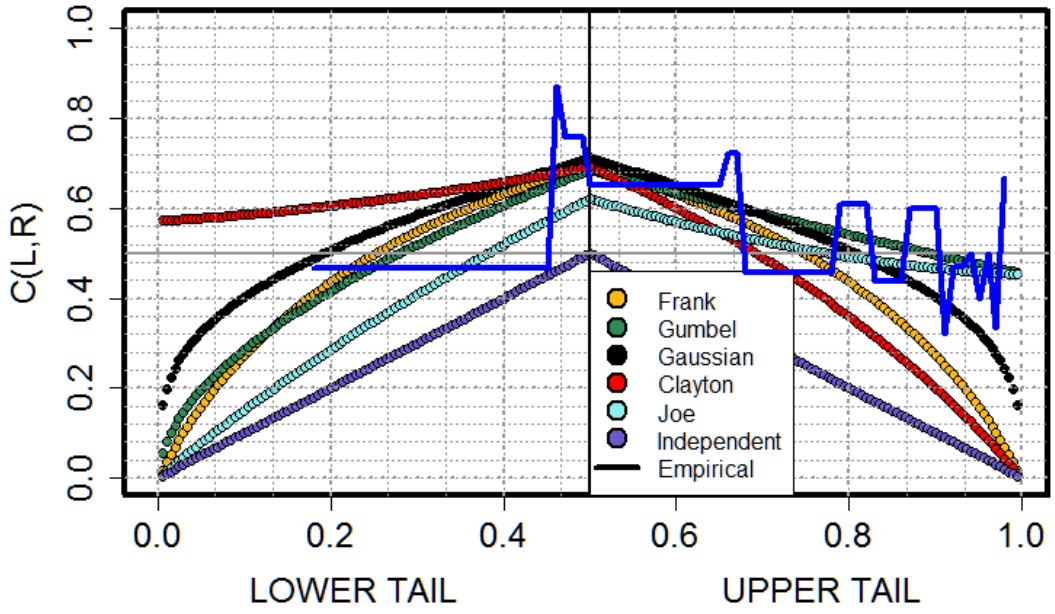


Figure (D.4.3-3). Tail dependence plot.