A MULTIVARIATE ANALYSIS OF FREEWAY SPEED AND HEADWAY DATA

A Dissertation

by

YAJIE ZOU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Yunlong Zhang
Committee Members,	Dominique Lord
	Xiubin Wang
	Faming Liang
Head of Department,	Robin Autenrieth

December 2013

Major Subject: Civil Engineering

ABSTRACT

The knowledge of speed and headway distributions is essential in microscopic traffic flow studies because speed and headway are both fundamental microscopic characteristics of traffic flow. For microscopic simulation models, one key process is the generation of entry vehicle speeds and vehicle arrival times. It is helpful to find desirable mathematical distributions to model individual speed and headway values, because the individual vehicle speed and arrival time in microscopic simulations are usually generated based on some form of mathematical models. Traditionally, distributions for speed and headway are investigated separately and independent of each other. However, this traditional approach ignores the possible dependence between speed and headway.

To address this issue, the dissertation presents two different methodologies to construct bivariate distributions to describe the characteristics of speed and headway. Based on the investigation of freeway speed and headway data measured from the loop detector data on IH-35 in Austin, it is shown that there exists a weak dependence between speed and headway and the correlation structure can vary depending on the traffic condition.

The dissertation first proposes skew-t mixture models to capture the heterogeneity in speed distribution. Finite mixture of skew-t distributions can significantly improve the goodness of fit of speed data. To develop a bivariate distribution to capture the dependence and describe the characteristics of speed and headway, finite mixtures of

multivariate skew-t distributions are applied to the 24-hour speed and headway data. The bivariate skew-t mixture model can provide a satisfactory fit to the multimodal speed and headway distribution and this modeling approach can accommodate the varying correlation structure between speed and headway.

To avoid the restriction of the bivariate skew-t distributions that individual behavior of speed and headway is described by the same univariate distributions, this research proposes copulas as an alternative method for constructing the multivariate distribution of traffic variables. Copula models can adequately represent the multivariate distributions of microscopic traffic data and accurately reproduce the dependence structure revealed by the speed and headway observations. This dissertation compares the advantages and disadvantages of copula models and finite mixtures of multivariate distributions. Overall, the proposed methodologies in this dissertation can be used to generate more accurate vehicle speeds and vehicle arrival times by considering their dependence on each other when developing microscopic traffic simulation models.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation and gratitude to my advisor, Dr. Yunlong Zhang, who provided me persistent guidance, support and help during my graduate study. Dr. Zhang is rich intellectually and can always come up with brilliant suggestions for conducting scientific research and conveying scientific results. He also encourages me to make independent thinking and gives me the freedom to explore on my own. His deep care and continued advice, both academically and personally, have made this work possible.

I would also like to thank my committee members, Dr. Dominique Lord, Dr. Bruce Wang, and Dr. Faming Liang for their time and suggestions of this dissertation. I have substantially benefitted from their classes during my Ph.D. study. Special thanks are given to Dr. Lord. I am grateful for his continuous guidance and detailed comments on my research work.

Also, many thanks to my friends at Texas A&M and fellow students in the transportation engineering division for the friendships and happiness they brought to me.

Last but not least, I would like to express my heart-felt gratitude to my wife Zhaoru Zhang and my parents for their unconditional support.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Speed is a fundamental measure of traffic performance of a highway system (May, 1990). Most analytical and simulation models of traffic either produce speed as an output or use speed as an input for travel time, delay, and level of service determination (Park et al., 2010). It is desirable to find an appropriate mathematical distribution to describe the measured speeds, because in some microscopic simulations the individual vehicle speed needs to be determined according to some form of mathematical model during vehicle generation (Park et al., 2010).

Headway is an important flow characteristic and headway distribution has applications in capacity estimation, driver behavior studies and safety analysis (May, 1990). The distribution of headway determines the requirement and the opportunity for passing, merging, and crossing (May, 1990). The headway distribution under capacity-flow conditions is also a primarily factor in determining the capacity of systems. Moreover, a key component in many microscopic simulation models is to generate entry vehicle headway in the simulation process. To generate accurate vehicle arrival times to the simulated network, it is necessary to use appropriate mathematical distributions to model headway.

As described above, the knowledge of speed and headway is necessary because these variables are fundamental measures of traffic performance of a highway system. Therefore, developing reliable and innovative analytical techniques for analyzing these variables is very important. The primary goal of this research is to develop some new methodologies for the analysis of microscopic freeway speed and headway data.

1.1 Statement of the Problem

This dissertation consists of three parts. The first part concerns the heterogeneity problem in freeway vehicle speed data. If the characteristics of speed data are homogeneous, speed can be generally modeled by normal, log-normal and gamma distributions. However, if the speed data exhibit excess skewness and bimodality (or heterogeneity), unimodal distribution function does not give a satisfactory fit. Thus, the mixture model (composite model) has been considered by May (1990) for traffic stream that consists of two classes of vehicles or drivers. So far, the mixture models used in previous studies to fit bimodal distribution of speed data considered normal density as the specified component; therefore, it is useful to investigate other types of component density for the finite mixture model.

The second and third parts concern the dependence between freeway speed and headway data. Traditionally, the dependence between speed and headway is ignored in the microscopic simulation models. As a result, the same headway distribution may be assumed for different speed levels and this assumption neglects the possible variability

of headway distribution across speed values. Moreover, a number of developed microscopic simulation models generate vehicle speeds and vehicle arrival times as independent inputs to the simulation process. Up to date, only a few studies have been directed at exploring the dependence between speed and headway. Considering the potential dependence between speed and headway, it is useful to construct bivariate distribution models to describe the characteristics of speed and headway. Compared with one dimensional statistical models representing speed or headway separately, bivariate distributions have the advantage that the possible correlation between speed and headway is taken into consideration. Given this advantage, it is necessary to construct bivariate distributions to improve the accuracy or validity of microscopic simulation models.

1.2 Research Objectives

The primary goal of this research is to develop new methodologies for analyzing the characteristics of speed and headway. To accomplish this goal, following objectives are planned to be addressed in this research.

1. To address the heterogeneity problem in freeway vehicle speed data, we apply skew-normal and skew-t mixture models to capture excess skewness, kurtosis and bimodality present in speed distribution. Skew-normal and skew-t distributions are known for their flexibility, allowing for heavy tails, high degree of kurtosis and asymmetry. To investigate the applicability of mixture models with skew-normal and

skew-t component density, we fit a 24-hour speed data collected on IH-35 using skew-normal and skew-t mixture models with the Expectation Maximization type algorithm.

2. To construct bivariate distribution of speed and headway, we examine the dependence structure between the two variables. Three correlation coefficients (i.e., Pearson correlation coefficient, Spearman's rho and Kendall's tau) are used to evaluate the dependence between speed and headway.

3. To develop a bivariate distribution for capturing the dependence and describing the characteristics of speed and headway simultaneously, finite mixtures of multivariate skew-t distributions are proposed. Finite mixtures of multivariate skew-t distributions have shown to be useful in modeling heterogeneous data with asymmetric and heavy tail behavior. In addition to the multivariate skew-t distribution, the multivariate normal and multivariate skew-normal distributions are also considered as the component density.

4. To avoid the restriction of the multivariate skew-t distributions that the individual behavior of the two variables is described by the same univariate distribution (i.e., skew-t distributions), copula models are proposed as an alternative method for constructing the multivariate distribution of traffic variables. Since vehicle type plays a role in the congested traffic condition, when constructing the multivariate distribution of traffic variables, vehicle length is used as a surrogate. The applicability of different families of copulas to traffic variables (speed, headway and vehicle length) is investigated and some recommendations are made.

1.3 Outline of the Dissertation

The rest of this dissertation is organized as follows:

Chapter II overviews various mathematical models that have been used for describing speed and headway distributions. Some studies that focused on the dependence between speed and headway are also discussed.

Chapter III provides the characteristics of the traffic dataset used throughout in the dissertation. A preliminary analysis is conducted to investigate the dependence structure between speed and headway.

Chapter IV applies skew-t mixture models to fit freeway speed data. This chapter shows that finite mixture of skew-t distributions can significantly improve the goodness of fit of speed data and better account for heterogeneity in the data.

Chapter V explores the applicability of the finite mixtures of multivariate distributions to address the heterogeneity problem in speed and headway data. This chapter shows that the bivariate skew-t mixture model can provide a satisfactory fit to the speed and headway data. This modeling approach can accommodate the varying correlation coefficient.

Chapter VI documents the application of copulas for constructing the multivariate distribution of traffic variables (speed, headway and vehicle length). This chapter

compares the advantages and disadvantages of copula models and finite mixtures of multivariate distributions.

Chapter VII summarizes the major results of in this research. General conclusions and recommendations for future research are presented.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

This chapter first provides a review of mathematical models for speed and headway. Specifically, different speed and headway distributions proposed in the past studies are introduced. Then, we discuss some research focused on the dependence between speed and headway.

2.2 Speed Distributions

Previously, normal, log-normal and other forms of distribution have been used to fit freeway speed data. Leong (1968) and McLean (1979) proposed that speed data approximately follow a normal distribution when flow rate is light. Haight and Mosher (1962) showed that the log-normal distribution is proper for speed data. Gerlough and Huber (1976) and Haight (1965) have used normal, log-normal and gamma distributions to model vehicular speed. Compared with normal distribution, log-normal and gamma distributions have the capacity to accommodate the right skewness and eliminate negative speed values generated by normal distribution. If the speed data exhibit excess skewness and bimodality, unimodal distribution function does not give a satisfactory fit; thus, several researchers used the mixture model to fit the distribution of speed. When the traffic stream consists of two vehicle types, the composite distribution has been proposed by May (1990). He also suggested that the vehicle speeds for subpopulations

follow normal or lognormal distributions. Dey et al. (2006) introduced a new parameter, spread ratio to predict the shape of the speed curve. He stated that the bimodal speed distribution curve consists of a mixture of two-speed fractions, lower fraction and upper fraction. Ko and Guensler (2005) did a similar study by characterizing the speed data with two different normal components, one for congested and the other for non-congested speeds. The congestion characteristics can be identified based on the speed distribution. Recently, Park et al. (2010) explored the distribution of 24-hour speed data with a g-component normal mixture model. Jun (2010) investigated traffic congestion trends by speed patterns during holiday travel periods using the normal mixture model.

2.3 Headway Distributions

Many headway models have been proposed and these models can be classified into two types: single distribution models and mixed models. For single distribution models, exponential (Cowan, 1975), normal, gamma, lognormal and log-logistic distributions (Yin et al., 2009) have been studied to model headway. The representatives of mixed models are Cowan M3 model (Luttinen, 1999), M4 model (Hoogendoorn and Bovy, 1998), the generalized queuing model and the semi-Poisson model (Wasielewski, 1979). Zhang et al. (2007) performed a comprehensive study of the performance of typical headway models using the headway data recorded from general-purpose lanes.

2.4 Dependence between Speed and Headway

There have been some studies that focused on the dependence between speed and headway. Luttinen (1992) found out that speed limit and road category have a considerable effect on the statistical properties of vehicle headways. WINSUM and Heino (1996) investigated the time headway and braking response during car-following. Taieb-Maimon and Shinar (2001) conducted a study to investigate drivers' following headways in car-following situation and the results showed that drivers adjusted the distance headways in relation to speed. Dey and Chandra (2009) proposed two statistical distributions for modeling the gap and headway in the steady car-following state. Brackstone et al. (2009) found that there is a limited dependence of following headway on speed and the most successful relationship fit of headway and speed is an inverse relationship. Yin et al. (2009) also studied the dependence of headway distributions on the traffic condition (speed pattern) and concluded that different headway models should be used for distinct traffic conditions (speed patterns).

2.5 Summary

From the above discussion, there are several current issues existing in modeling the speed and headway data. First, when modeling multimodal distribution of speed data, the mixture models used in previous studies extensively considered normal density as the specified component; therefore, other types of component density were not fully investigated. Second, considering the possible dependence between speed and headway,

there were very few studies focusing on constructing bivariate distribution models to describe speed and headway simultaneously.

CHAPTER III

DATA INTRODUCTION AND PRELIMINARY ANALYSIS

3.1 Introduction

As discussed in Chapter I, the main objective of this dissertation is to develop new methodologies for analyzing the characteristics of freeway speed and headway data. The traffic data analyzed in this dissertation are the microscopic traffic variables (i.e., individual speed and headway observations) measured from the loop detector data. The study site is on IH-35 in Austin, Texas. This chapter introduces the characteristics of the traffic dataset which is used throughout in the dissertation. A preliminary analysis is conducted to investigate the dependence structure between observed speed and headway data.

3.2 Data Description

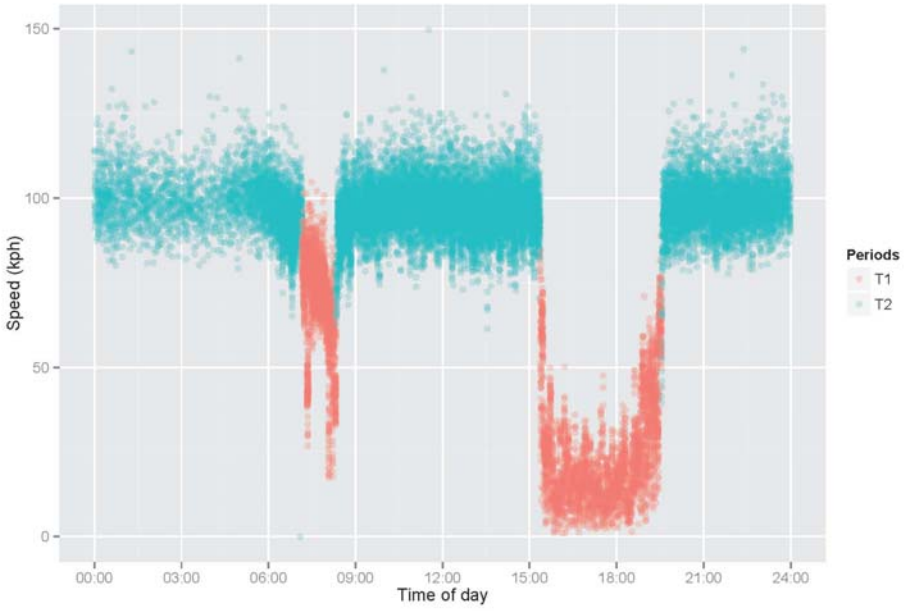
The dataset was collected at a location on IH-35. IH-35 has four lanes in the southbound direction and the free flow speed is 60 mile/hour (or 96.56 kilometer/hour) for all types of vehicles. Due to the heavy traffic demand and a large volume of heavy vehicles, the data collection site is typically congested during the morning and afternoon peak hours. The detector records vehicle arrival time, presence time, speed, length, and classification for each individual vehicle (Ye et al., 2006). This dataset was analyzed in some previous studies (Ye and Zhang, 2009). The data have 27920 vehicles with recorded speed values, arrival times and vehicle lengths in a 24-hour period (from 00:00 to 24:00, December 11,

2004), including 24011 (86%) passenger vehicles and 3909 (14%) heavy vehicles. For this dataset, the headway value between two consecutive vehicles is the elapsed time between the arrivals of a pair of vehicles. The arrival times were recorded in second (s); the observed speeds were recorded in meter/second; and the vehicle lengths were recorded in meter (m). To compare the result of this work with some previous studies, we convert the meter/second to kilometer/hour (kph). We also assume that 24-hour period (T) consists of two time periods: the peak time period (T1) which contains two sub-periods 07:10-08:20 and 15:22-19:33; while the off peak period (T2) includes two sub-periods 08:20-15:22 and 19:33-07:10.

3.3 Preliminary Analysis

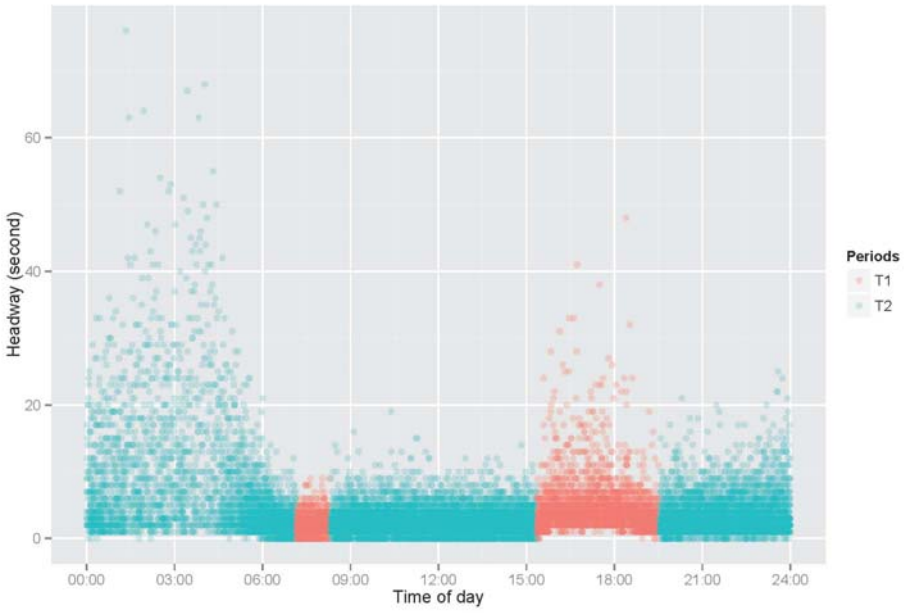
Figure 3.1 (a), (b) and (c) display the scatter plots of speed, headway and vehicle length by time of day for each time period. Because of large samples in the dataset, semi-transparent points are used to alleviate some of the over-plotting in Figure 3.1. Figure 3.1 (c) indicates that the observed vehicles seem to consist of two sub-populations: one at about 5 meters, representing passenger vehicles, and the other at about 22 meters, representing trucks and buses. Previously, Zhang et al. (2008) estimated large truck volume using loop detector data collected from IH-35, and they classified vehicles into two categories: short vehicles (smaller than 12.2 m (40 feet)) and long vehicles (larger than or equal to 12.2 m (40 feet)). In order to see the changing pattern of vehicle composition over the time, we calculate the hourly percentage of long vehicles (greater than or equal to 12.2 m), which is shown in Figure 3.1 (d). It can be observed that the

proportion of long vehicles is relatively high between 00:00 and 6:00 compared with other time periods of the day.

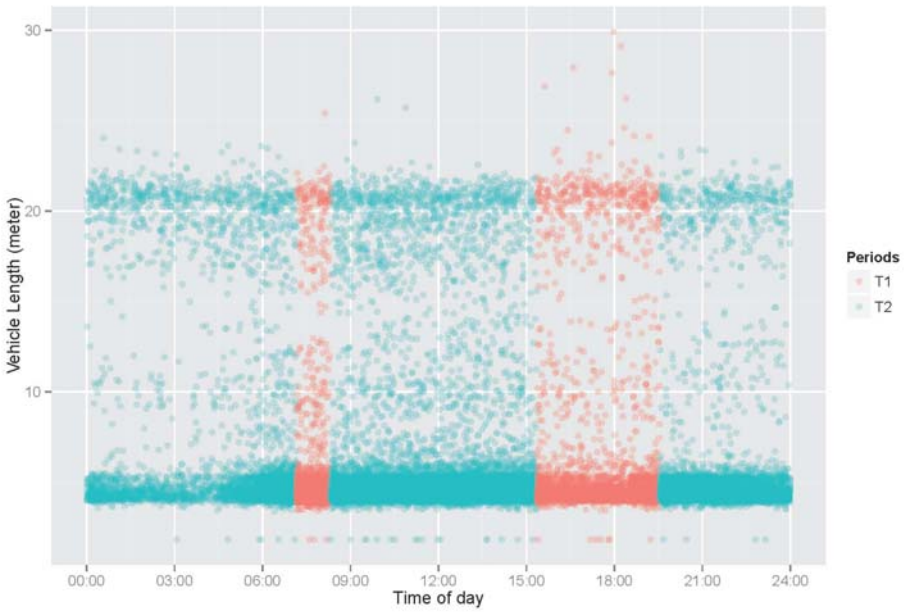


(a)

Figure 3.1 (a) speed scatter plots by time of day; (b) headway scatter plots by time of day; (c) vehicle length scatter plots by time of day; (d) hourly percentage of long vehicles by time of day.

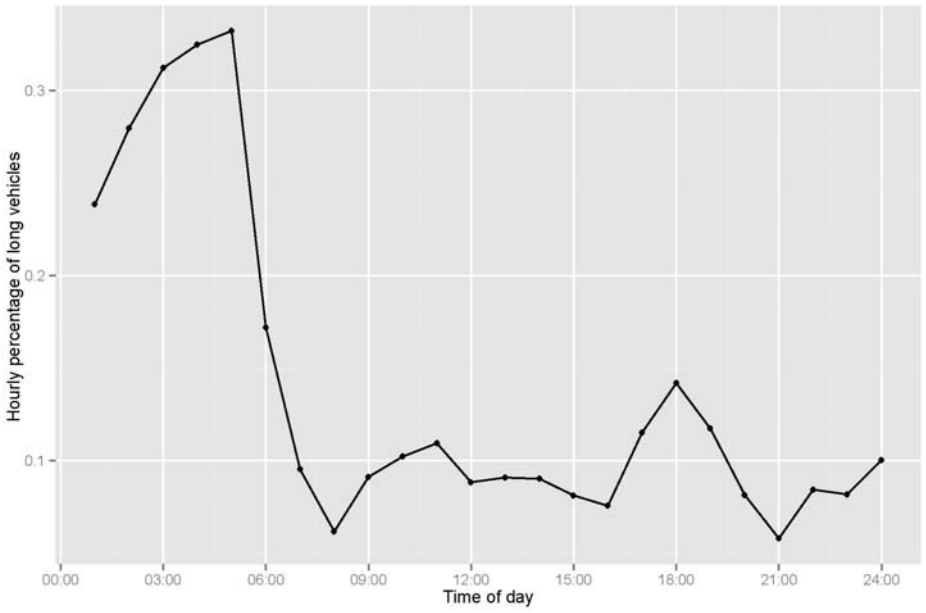


(b)



(c)

Figure 3.1 Continued



(d)

Figure 3.1 Continued

From Figure 3.1 (a), we can see that the speed data exhibit heterogeneity and the main cause for this heterogeneity is different traffic flow conditions over the 24-hour period. Since the characteristics of speed data are heterogeneous, the mixture models are used to capture bimodality present in speed distribution. Then, we examine the correlation between speed and headway. Since the 24-hour traffic data in the study consist of distinct traffic flow conditions, it is useful to evaluate the dependence between vehicle speed and headway under different traffic conditions. As discussed above, we divided the 24-hour traffic data into two time periods (i.e., the peak period T1 and the off-peak period T2) based on corresponding traffic conditions. For each time period, three correlation coefficients are used to evaluate the dependence. These three measures of

dependence are Pearson correlation coefficient (PCC), Spearman's tau (SCC), and Kendall's pho (KCC). The summary statistics of speed and headway for different time periods are given in Table 3.1.

Table 3.1 Summary statistics of speed and headway for different time periods

	T (24 hours)		T1 (07:10-08:20 and 15:22-19:33)		T2 (08:20-15:22 and 19:33-07:10)	
	Speed	Headway	Speed	Headway	Speed	Headway
Min.	0	0 [a]	1.01	0	0	0
1st Quantile	84.74	1	18.22	2	92.38	1
Median	94.57	2	37.76	2	97.09	2
Mean	85.3	3.1	42.71	3.15	97.24	3.08
3rd Quantile	100.4	3	68.57	4	101.95	3
Max.	149.69	76	104.72	48	149.69	76
Number of vehicles	27919		6114		21805	
PCC	-0.054		-0.469		0.116	
KCC	0.003		-0.488		0.135	
SCC	0.011		-0.635		0.186	

Note: [a] Headway values are less than 0.5s.

PCC measures the linear relationship between two continuous variables. It is defined as the ratio of the covariance of the two variables to the product of their respective standard deviations:

$$\mathrm{PCC} = \frac{Cov(x, y)}{\sigma_x \sigma_y} \tag{3.1}$$

where σ_x and σ_y are the standard deviations of variables x and y.

SCC is a rank-based version of the PCC and it can be computed as:

$$\mathrm{SCC} = \frac{\sum_{i=1}^{n} (rank(x_i) - \overline{rank(x)})(rank(y_i) - \overline{rank(y)})}{\sqrt{\sum_{i=1}^{n} (rank(x_i) - \overline{rank(x)})^2 \sum_{i=1}^{n} (rank(y_i) - \overline{rank(y)})^2}} \tag{3.2}$$

where $rank(x_i)$ and $rank(y_i)$ are the ranks of the observation x_i and y_i in the sample.

Similar to SCC, KCC is designed to capture the association between two measured quantities. KCC quantifies the discrepancy between the number of concordant and discordant pairs. Its estimate can be expressed as follows:

$$\mathrm{KCC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{sgn}(x_i - x_j)\,\mathrm{sgn}(y_i - y_j)}{\frac{1}{2} n(n-1)} \tag{3.3}$$

where $\mathrm{sgn}(x_i - x_j) = \begin{cases} 1 & \text{if } (x_i - x_j) > 0 \\ 0 & \text{if } (x_i - x_j) = 0 \\ -1 & \text{if } (x_i - x_j) < 0 \end{cases}$ and $\mathrm{sgn}(y_i - y_j) = \begin{cases} 1 & \text{if } (y_i - y_j) > 0 \\ 0 & \text{if } (y_i - y_j) = 0 \\ -1 & \text{if } (y_i - y_j) < 0 \end{cases}$.

Note that the PCC, KCC, and SCC are -0.469, -0.488 and -0.635 between speed and headway for peak period T1, suggesting a moderate inverse relationship between these two traffic variables. Since speed and headway values in peak period T1 were observed under congested traffic conditions, it is reasonable to consider most of the headway values in time period T1 as following headways. From Figure 3.2, it is observed that headway increases as speed decreases, and the relationship can be split into two regimes. The time headway is approximately stable when speed is above 20 kph in the first regime. In the second regime when speed is below 20 kph, the time headway increases significantly as speed decreases. The findings from Figure 3.2 are consistent with the results reported in a study conducted by Brackstone et al. (2009). In their study, it is shown that there is a limited dependence of following headway on speed: the most successful relationship fit of headway and speed is an inverse relationship. Interestingly, KCC is 0.135 between speed and headway for off-peak period T2, indicating a positive dependence. This is reasonable because as headway values become larger during the off peak period, fewer vehicles are on the road and it is expected to see that vehicle speeds increase accordingly.

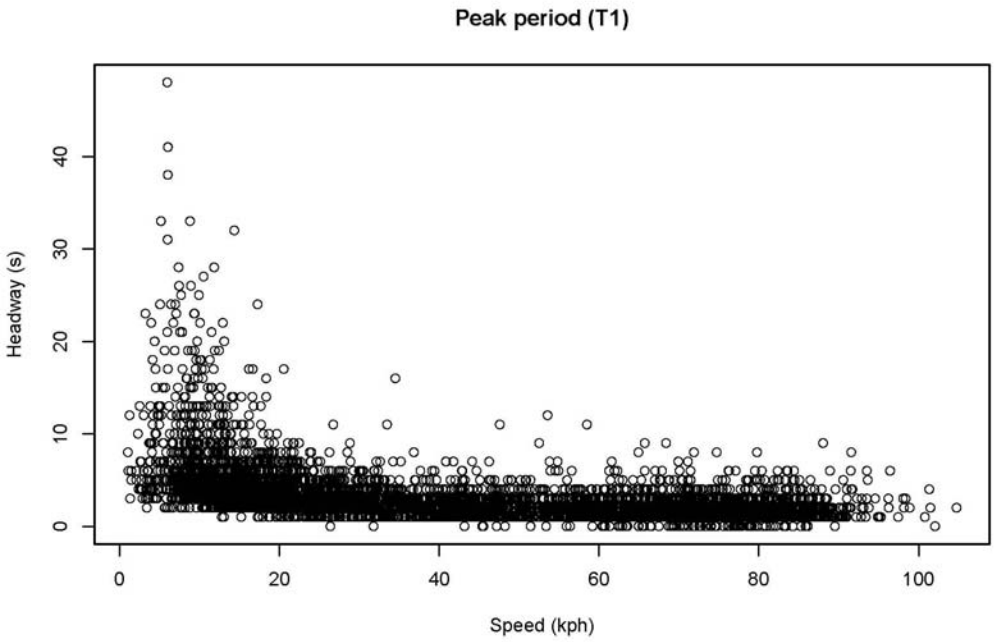


Figure 3.2 Scatter plot of speed and headway for peak period (T1).

3.4 Summary

This chapter described the characteristics of traffic data collected on IH-35. As shown in Figure 3.1 (a), the speed data are heterogeneous and to capture the bimodality present in the speed distribution, Chapter IV proposes skew-t mixture models to fit freeway speed data. Besides, the data analysis indicates that the two microscopic traffic variables (speed and headway) are correlated under different traffic conditions, and the correlation structure tends to vary depending on the traffic condition. Thus, in order to construct bivariate distribution of speed and headway, two different methodologies (i.e., finite mixtures of multivariate skew-t distributions and copula models) are proposed in Chapters V and VI, respectively.

CHAPTER IV

METHODOLOGY I: MIXTURE MODELING OF FREEWAY SPEED DATA[1]

4.1 Introduction

An appropriate mathematical distribution can help describing speed characteristics and is also useful for developing and validating microscopic traffic simulation models. To accommodate the heterogeneity in speed data, the mixture models used in previous studies extensively considered normal density as the specified component; therefore, other types of component density were not fully investigated. To capture excess skewness, kurtosis and bimodality present in speed distribution, we propose skew-normal and skew-t mixture models to fit freeway speed data. This chapter shows that finite mixture of skew-t distributions can significantly improve the goodness of fit of speed data and better account for heterogeneity in the data.

4.2 Finite Mixture Models

In this chapter, it is assumed that the speed data are independent and identically distributed (i.i.d.) realizations from a random variable which follows either a mixture of g-component normal, skew-normal or skew-t mixture model. The mixture model is

---

[1] Reprinted with permission from "Use of skew-normal and skew-t distributions for mixture modeling of freeway speed data" by ZOU, Y., & ZHANG, Y., 2011. Transportation Research Record, 2260, 67-75, Copyright [2011] by the Transportation Research Board. None of this material may be presented to imply endorsement by TRB of a product, method, practice, or policy.

widely used in modeling bimodal speed distribution to account for the heterogeneity. The normal, skew-normal and skew-t mixture models are briefly introduced in this section:

The normal mixture model for the vehicle speed has the following probability density function:

$$f(x \mid w_k, \xi_k, \sigma_k^2) = \sum_{k=1}^{N} w_k NL(x \mid \xi_k, \sigma_k^2) \tag{4.1}$$

$$NL(x \mid \xi_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{(x-\xi_k)^2}{2\sigma_k^2}) \tag{4.2}$$

The expectation and variance of a normal distribution can be written as:

$$E(x) = \xi_k \tag{4.3}$$

$$Var(x) = \sigma_k^2 \tag{4.4}$$

where N is the number of components, w_k is the weight of component k, with $1 > w_k > 0$ and $\sum_{k=1}^{N} w_k = 1$, ξ_k is the location parameter, σ_k^2 is the scale parameter, and $NL(x \mid \xi_k, \sigma_k^2)$ is the normal density function with mean ξ_k and variance σ_k^2.

The skew-normal distribution was first developed by Azzalini (1985). The probability density function for the skew-normal mixture model is given by:

$$f(x \mid w_k, \xi_k, \sigma_k^2, \lambda_k) = \sum_{k=1}^{N} w_k SN(x \mid \xi_k, \sigma_k^2, \lambda_k) \tag{4.5}$$

$$SN(x \mid \xi_k, \sigma_k^2, \lambda_k) = \frac{2}{\sigma_k} \phi\left(\frac{x-\xi_k}{\sigma_k}\right) \Phi\left(\lambda_k \frac{x-\xi_k}{\sigma_k}\right) \tag{4.6}$$

The expectation and variance of a skew-normal distribution are given by

$$E(x) = \xi_k + \sigma\delta_k \sqrt{\frac{2}{\pi}} \tag{4.7}$$

$$Var(x) = \sigma_k^2 \left(1 - \frac{2\delta_k^2}{\pi}\right) \tag{4.8}$$

where $\delta_k = \frac{\lambda_k}{\sqrt{1+\lambda_k^2}}$, λ_k is the skewness parameter, $\phi(\cdot)$ and $\Phi(\cdot)$ are, the standard normal density and cumulative distribution function, and $SN(x \mid \xi_k, \sigma_k^2, \lambda_k)$ is the skew-normal density function. The mean and variance of $SN(x \mid \xi_k, \sigma_k^2, \lambda_k)$ are given in equations (4.7) and (4.8), respectively.

It can be shown that the excess kurtosis of a skew-normal distribution is limited to the interval [0, 0.8692]. Later, the skew-t distribution was introduced by Azzalini and Capitanio (2003) to allow for a higher degree of kurtosis. The skew-t mixture model can be written as follows:

$$f(y \mid w_k, \xi_k, \sigma_k^2, \lambda_k, v) = \sum_{k=1}^{N} w_k ST(y \mid \xi_k, \sigma_k^2, \lambda_k, v) \tag{4.9}$$

$$ST(y \mid \xi_k, \sigma_k^2, \lambda_k, \nu) = \frac{2}{\sigma_k} t_\nu(x_y) T_{\nu+1}\left(\lambda_k x_y \sqrt{\frac{\nu+1}{\nu + x_y^2}}\right) \tag{4.10}$$

where ν is the degrees of freedom, $x_y = (y - \xi_k)/\sigma_k$, t_ν and T_ν represent the standard Student-t density and cumulative function with ν degrees of freedom, and $ST(y \mid \xi_k, \sigma_k^2, \lambda_k, \nu)$ is the skew-t density function. Also, it can be shown that the skew-t distribution converges to a skew-normal distribution when $\nu \to \infty$ (ν tends to infinity).

4.3 Model Estimation Method

There are various methods available for estimating a mixture model. The method of moments was first used by Pearson in the early days of mixture modeling. The maximum likelihood estimation with Expectation Maximization (EM) algorithm and Bayesian estimation become the most widely applied methods when large calculations can be easily done by powerful computers. Assuming the number of components is known, Bayesian approach can be implemented with data augmentation and Markov Chain Monte Carlo (MCMC) estimation procedure using Gibbs sampling techniques (Zou et al, 2012). However, one of the main drawbacks of MCMC procedures is that they are generally computationally demanding, and it can be difficult to diagnose convergence (Zou et al, 2012). Furthermore, the label switching is another difficulty and has to be addressed explicitly when using a Bayesian approach to conduct parameter estimation and clustering (Frühwirth-Schnatter, 2006).

Since the label switching is of no concern for maximum likelihood estimation, the maximum likelihood method is adopted for estimation of finite mixture of skew-normal and skew-t distributions in this study. The EM algorithm was introduced by Dempster et al. (1977) and there are two extensions of it: the Expectation/Conditional Maximization Either (ECME) and the Expectation/Conditional Maximization (ECM) algorithms. Among the three algorithms, the ECM algorithm converges more slowly than the EM algorithm, but consumes less processing time in computer. The ECME algorithm has the greatest speed of convergence as well as the least processing time; moreover, it preserves the stability with monotone convergence. Thus, the ECME algorithm is chosen for the estimation of the parameters here.

4.4 Modeling Results

We apply normal, skew-normal and skew-t mixture models with an increasing number of components (g =2,…,6) to the 24-hour speed data described in Chapter III. The ECME algorithm is coded and run until the convergence maximum error 0.0000001 is satisfied or until the maximum number of iterations 3000 is reached. A common problem with this method is that the EM type algorithm may lead to a local maximum and one feasible solution to find the global maximum is to try many different initial values. Therefore, the procedure described by Basso et al. (2010) is adopted to ensure that initial values are not far from the real parameter values.

4.4.1 Determination of optimal model

To select the most appropriate model from normal, skew-normal and skew-t mixture models, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the Integrated Completed Likelihood Criterion (ICL) are computed for each mixture model. AIC and BIC have the same form $-2LL + \gamma c_n$, where LL is the log-likelihood value, γ is the number of free parameters to be estimated and c_n is the penalty term with a positive value.

The value of c_n is defined depending on the selected criterion. For AIC and BIC, c_n equals 2 and $\log(n)$ respectively, where n is the number of observations. The ICL criterion approximated from a BIC-like approximation is defined as $-2LL^* + \gamma \log(n)$, where LL^* is the integrated log-likelihood. It is known that BIC is more conservative than AIC. In the density estimation context, BIC is a reliable tool for comparing mixture models. When choosing the form of the model, using BIC as the criterion usually results in a good fit of data. If the finite mixture model is correctly specified, BIC is known to be consistent. On the other hand, if the concern of mixture modeling is cluster analysis, ICL criterion is preferred over BIC when selecting the optimal number of components g, because BIC may overestimate the number of components (Biernacki et al., 2000). In particular, BIC is likely to be imprecise in identifying the correct size of the clusters when component densities of mixture model are not specified correctly. The ICL

criterion includes an additional entropy term which favors well-separated clusters (Biernacki et al., 2000).

Bold values in Table 4.1 report the smallest AIC, BIC among three mixture models. Smaller AIC and BIC values indicate a better overall fit. Based on the results, the skew-t mixture model is selected as the best one for g = 2, 3, 5, 6. For g = 4, the skew-normal mixture model is slightly better than the skew-t mixture model in terms of AIC and BIC values. Upon comparison of three mixture models, we find that the skew-normal and skew-t mixture models both show a much better fitting result than the normal mixture model; the skew-t mixture model has the smallest AIC and BIC values except when g equals 4. The computation times for each model are shown in Table 4.1. Compared with the normal mixture model, the skew-normal mixture model can significantly improve the goodness of fit of speed data while the increase in computational effort is not remarkable. Given this advantage, the skew-normal mixture model can be used as an alternative to the skew-t mixture model if the computation time is limited. And the skew-t mixture model can achieve the best fitting result at the cost of more computation time.

Another important criterion considered for model assessment is the Kolmogorov-Smirnov's (K-S) goodness of fit test (Lin et al., 2007). We performed K-S tests to validate the above three mixture models. The statistics D and p-value for K-S tests are summarized in Table 4.2. Note that in a K-S test, given a sufficiently large sample, a small and non-notable statistics D can be found to be statistically significant. For normal,

skew-normal and skew-t mixture models, normal and skew-normal model with 2 components are rejected and none of skew-t mixture models is rejected when the significance level is 0.01. Thus, it also suggests that speed data can be better described by a mixture of skew-t distributions.

In summary, the skew-t mixture model outperforms the other two mixture models based on AIC, BIC and K-S test results. We select the skew-t mixture model as the best one and use it to determine the number of components. The parameter estimation results for the skew-t mixture distribution are provided in Table 4.3.

Table 4.1 Computed AIC, BIC and ICL values for three mixture models

g = 2	Normal	Skew-normal	Skew-t
AIC	232936.8	230936	**230223.5**
BIC	232978	230977.1	**230264.7**
ICL	234836.4	233345.5	**231732.3**
Time*	1 min	4 mins	45 mins
g = 3	Normal	Skew-normal	Skew-t
AIC	230254.7	229819.3	**229811.7**
BIC	230320.6	229885.2	**229877.6**
ICL	235846.5	242316.9	**235082.6**
Time*	1 min	6 mins	63 mins

Table 4.1 Continued

g = 4	Normal	Skew-normal	Skew-t
AIC	229921.4	**229801.9**	229802
BIC	230012	**229892.5**	229892.6
ICL	**239894**	256410.1	250663.8
Time*	4 mins	8 mins	363 mins
g = 5	Normal	Skew-normal	Skew-t
AIC	229836.3	229745	**229740.6**
BIC	229951.6	229860.4	**229855.9**
ICL	251178.7	**247112.5**	251844.7
Time*	8 mins	22 mins	438 mins
g = 6	Normal	Skew-normal	Skew-t
AIC	229809.1	229786.7	**229746**
BIC	229949.2	229926.7	**229886**
ICL	257663.9	**243317.1**	245020.3
Time*	18 mins	32 mins	518 mins

* These experiments were performed on a desktop with Core 2 Duo processor E8500 running at 3.16 GHz and 4 GB RAM.

Table 4.2 The K-S test results for three mixture models

No. of components	Normal		Skew-normal		Skew-t	
	D	p-value	D	p-value	D	p-value
g = 2	0.0275	0.0000	0.0220	0.0000	0.0146	0.0109
g = 3	0.0117	0.04242	0.0074	0.4796	0.0074	0.4825
g = 4	0.009	0.2055	0.0072	0.5016	0.0071	0.5141
g = 5	0.007	0.5038	0.0069	0.5444	0.0070	0.5256
g = 6	0.0067	0.5583	0.0073	0.4894	0.0067	0.5764

4.4.2 Selecting the number of components

It is quite a challenge to determine the optimal number of components in finite mixture models. Currently, available methods include reversible jump MCMC and model choice criteria. For skew-t mixture models, the implementation of reversible jump MCMC turns out to be very complicated and computation of marginal likelihoods remains an issue. Thus, we adopted the model choice criteria. As mentioned before, AIC tends to select too many components and BIC overrates the number of components if the component densities are misspecified. ICL criterion seems to provide a reliable estimate of g for real data (Biernacki et al., 2000). Thus, ICL values reported in Table 4.1 are used to determine the optimal number of components. Based on ICL criterion, g = 2 is chosen for the skew-t mixture model. Previously, Park et al. (2010) explored the data with a normal mixture model and selected the optimal number of components g = 4. To provide

further insight into the pattern of mixture, we fit the speed distribution with a 2-component skew-t mixture model and a 4-component normal mixture model.

The mixture density as well as each component-wise density for the 2-component skew-t and 4-component normal mixture distributions are displayed in Figure 4.1 and Figure 4.2, respectively. Based on the graphical visualization, both 2-component skew-t and 4-component normal mixture models fit the 24-hour speed distribution very well. However, as shown in these figures, the bimodality of the speed distribution suggests the presence of 2 different speed groups. One skew-t distribution can adequately capture the skewness and kurtosis present in one cluster; by contrast, two normal mixtures are needed to accommodate the skewness and kurtosis of one speed group. It is observed in Figure 4.1 that cluster 1 is composed of speed data from group 1 and cluster 2 consists of speed data from group 2. Since group 1 and group 2 represent distinct traffic flow characteristics, this verifies that traffic flow condition is the main cause for heterogeneity in this 24-hour speed data. On the other hand, no clear interpretation can be made regarding different flow conditions if a 4-component normal mixture model is used.

To summarize, the skew-t mixture model classified vehicle speed into 2 clusters. Component 1 (high speed cluster) includes vehicles in uncongested traffic condition and a large portion of vehicles in transition flow condition. Component 2 (low speed cluster) has a large variance and represents vehicles in congested traffic condition and a small portion of vehicles in transition flow condition.

Table 4.3 Parameter estimation results for the Skew-t mixture distribution

Component	Parameters	1	2	3	4	5	6
g=2	ξ	101.71	6.96				
	σ^2	79.01	491.72				
	λ	-1.07	8.06				
	nu*	3.59	3.59				
	η	0.85	0.15				
g=3	ξ	88.21	93.92	7.27			
	σ^2	298.74	55.04	363.48			
	λ	-1.08	0.72	6.09			
	nu*	9.33	9.33	9.33			
	η	0.14	0.73	0.13			
g=4	ξ	78.36	93.66	7.23	99.78		
	σ^2	254.36	85.60	375.22	90.90		
	λ	-2.33	1.92	6.09	-1.52		
	nu*	15.05	15.05	15.05	15.05		
	η	0.07	0.39	0.13	0.41		

Table 4.3 Continued

g=5	ξ	40.41	7.67	93.05	71.68	100.58	
	σ^2	41.52	294.79	99.96	88.97	100.36	
	λ	2.98	4.97	2.49	-0.50	-1.59	
	nu*	20.43	20.43	20.43	20.43	20.43	
	η	0.01	0.12	0.35	0.06	0.45	
g=6	ξ	7.85	93.35	99.59	91.89	38.55	70.93
	σ^2	270.45	46.18	95.12	83.40	943.53	36.68
	λ	4.52	1.05	2.05	-0.90	9.99	-1.20
	nu*	100.00	100.00	100.00	100.00	100.00	100.00
	η	0.12	0.54	0.11	0.17	0.05	0.02

* Kurtosis parameter

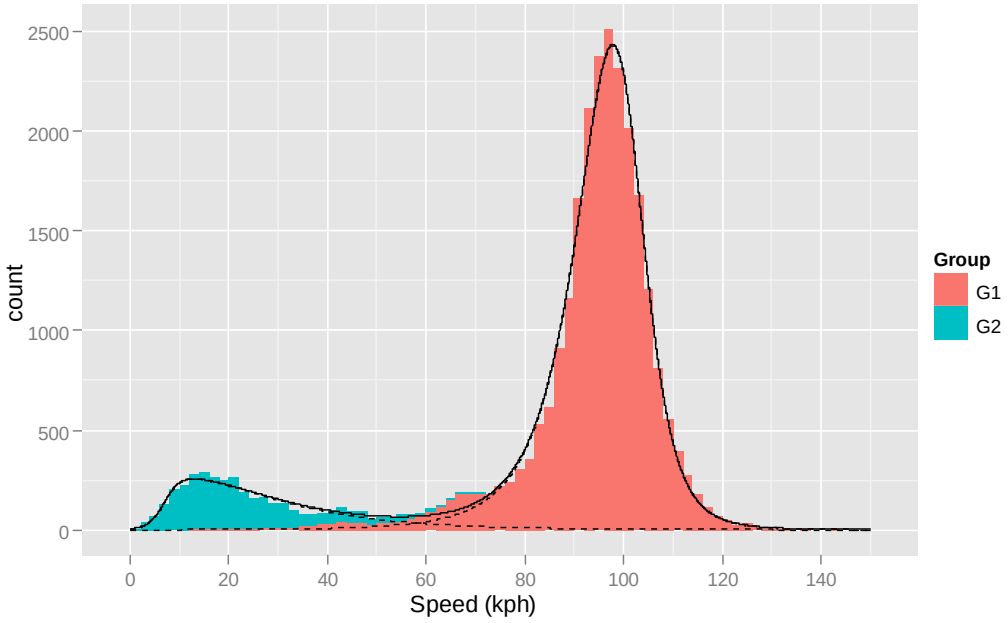


Figure 4.1 The fitted mixture model for 2-component skew-t distribution.

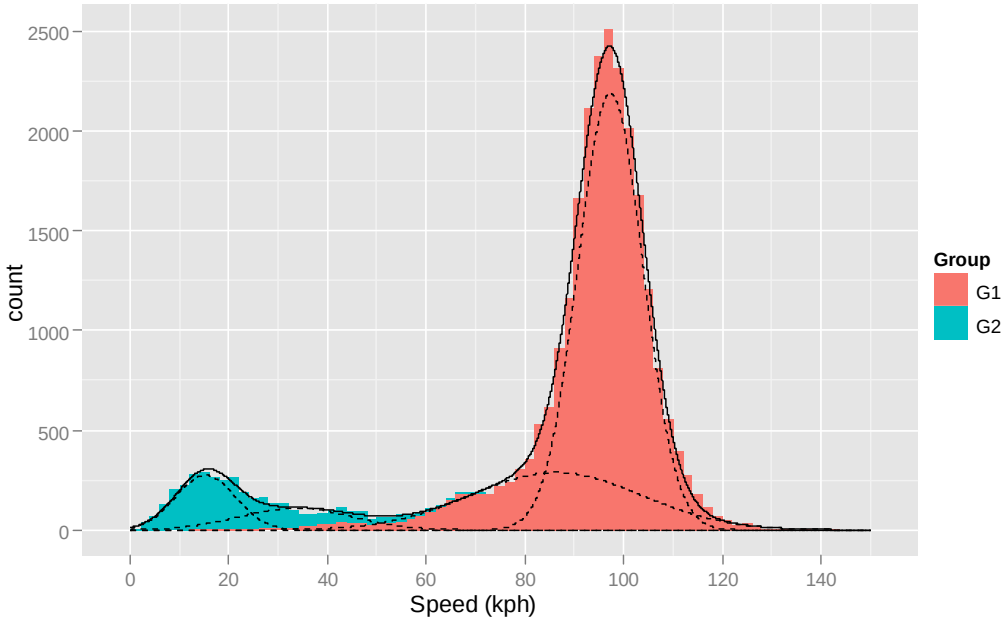


Figure 4.2 The mixture model for 4-component normal distribution.

4.5 Summary

This chapter has shown that skew-t distributions are useful for fitting the distribution of speed data. It is observed that for heterogeneous traffic flow condition, the flexibility of bimodal distribution causes problems when normal mixture models are used. The skew-t distributions are preferred component densities because they can capture skewness and excess kurtosis themselves. The finite mixture of skew-t distributions can significantly improve the goodness of fit of speed data.

CHAPTER V

METHODOLOGY II: MULTIVARIATE MIXTURE MODELING OF FREEWAY SPEED AND HEADWAY DATA

5.1 Introduction

To construct a bivariate distribution of speed and headway that can accommodate the heterogeneity in speed and headway data, finite mixtures of multivariate skew-t distributions are proposed in this study. Finite mixtures of multivariate skew-t distributions have shown to be useful in modeling heterogeneous data with asymmetric and heavy tail behavior (Lee and McLachlan, 2013). Besides the multivariate skew-t distribution, the multivariate normal and multivariate skew-normal distributions are also considered as the component density. This chapter shows that finite mixtures of multivariate skew-t distributions can provide a satisfactory fit to the speed and headway distribution.

5.2 Basic Assumptions

Drivers' speed and headway choices are jointly determined by some factors: driving-related factors (age, driver experience, alcohol level and so on); factors related to vehicle and road (roadway geometric configurations, vehicle types, etc.); and traffic or environment-related factors (traffic flows, vehicle composition, traffic control, etc.). Unfortunately, some factors (i.e., driving-related data) are usually not observable. The correlation structure between speed and headway are likely to be influenced by some

factors. Thus, it is reasonable to assume that speed and headway data with different combinations of factors (i.e., traffic conditions, etc.) can be divided into distinct sub-populations (the correlation structure between speed and headway is different across and similar within the sub-populations). In this study, it is assumed that the individual vehicle speed and headway are generated from a certain number of sub-populations.

5.3 Multivariate Distributions of Speed and Headway

5.3.1 Multivariate normal distribution

According to Tong (1990), the p-variate normal distribution $N_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, has the following density

$$N_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p \mid \boldsymbol{\Sigma} \mid}} \exp\left(-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right) \tag{5.1}$$

where $\mathbf{y}$ is the $p \times 1$ observation vector, $\boldsymbol{\mu}$ is the $p \times 1$ mean vector, $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix, and $\mid \boldsymbol{\Sigma} \mid$ is the determinant of $\boldsymbol{\Sigma}$. When $p = 1$, the density of the univariate normal distribution is defined as:

$$N_1(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-\mu)^2}{2\sigma^2}) \tag{5.2}$$

where σ^2 is the variance.

5.3.2 Multivariate skew-normal distribution

Different characterizations of the multivariate skew-normal and skew-t distributions have been developed in recent years (see Lee and McLachlan (2013) for an overview of

the various parameterizations of the multivariate skew-normal and skew-t distributions). The multivariate skew-normal distribution used in this research was developed by Azzalini and DallaValle (1996). The p-variate skew-normal distribution $SN_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, has the following density

$$SN_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}) = 2\phi_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\Phi(\boldsymbol{\lambda}^T \boldsymbol{\Sigma}^{-1/2}(\mathbf{y} - \boldsymbol{\mu})) \tag{5.3}$$

where $\boldsymbol{\lambda}$ is the $p \times 1$ shape parameter vector, $\boldsymbol{\lambda}^T$ denotes the transpose of $\boldsymbol{\lambda}$, $\boldsymbol{\Sigma}^{-1/2}$ is the root of $\boldsymbol{\Sigma}$, $\phi_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the density of the p-variate normal distribution $N_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and $\Phi(\bullet)$ is the cumulative distribution function of the standard univariate normal distribution. Note that when $\boldsymbol{\lambda}$=0, $SN_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$ reduces to the normal distribution $N_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$. For the univariate skew-normal distribution, the density of $SN_1(y \mid \mu, \sigma^2, \lambda)$ is given by

$$SN_1(y \mid \mu, \sigma^2, \lambda) = \frac{2}{\sigma}\phi_1\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right) \tag{5.4}$$

where μ is the location parameter, σ^2 is the scale parameter, and $\phi_1(\bullet)$ is the standard univariate normal density function.

5.3.3 Multivariate skew-t distribution

The multivariate skew-t distribution was first developed by Azzalini and Capitanio (2003). The p-variate skew-t distribution with ν degrees of freedom $ST_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$, has the following density

$$ST_p(\mathbf{y}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\lambda},\nu) = 2t_p(\mathbf{y}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)T\left(\sqrt{\frac{\nu+p}{\nu+d_{\boldsymbol{\Sigma}}(\mathbf{y},\boldsymbol{\mu})}}\boldsymbol{\lambda}^T\boldsymbol{\Sigma}^{-1/2}(\mathbf{y}-\boldsymbol{\mu})\,|\,\nu+p\right) \tag{5.5}$$

where $t_p(\bullet|\,\boldsymbol{\mu},\boldsymbol{\Sigma},\nu)$ stands for the density of the p-variate Student-t distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$ and ν degrees of freedom, $T(\bullet|\,\nu+p)$ is the cumulative distribution function of the standard univariate student-t distribution with $\nu+p$ degrees of freedom and $d_{\boldsymbol{\Sigma}}(\mathbf{y},\boldsymbol{\mu}) = (\mathbf{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})$. When $\nu\to+\infty$, the skew-t distribution converges to a skew-normal distribution. The density of the univariate skew-t distribution can be written as:

$$ST_1(y\,|\,\mu,\sigma^2,\lambda,\nu) = \frac{2}{\sigma}t_1(x_y\,|\,\nu)T\left(\lambda x_y\sqrt{\frac{\nu+1}{\nu+x_y^2}}\middle|\,\nu+1\right) \tag{5.6}$$

where $x_y = (y-\mu)/\sigma$, t_1 denotes the standard univariate Student-t density function.

5.3.4 Finite mixtures of multivariate distributions

The probability density function (PDF) of a g-component mixture of multivariate distributions is given by

$$f(\mathbf{y}\,|\,\boldsymbol{\Theta}) = \sum_{j=1}^{g} w_j\psi(\mathbf{y}\,|\,\boldsymbol{\theta}_j) \tag{5.7}$$

where w_j is the weight of component j, $w_j \geq 0$, $\sum_{j=1}^{g} w_j = 1$, $\boldsymbol{\Theta} = ((\boldsymbol{\theta}_1^T, w_1),...,(\boldsymbol{\theta}_g^T, w_g))^T$ is the vector of all parameters, $\boldsymbol{\theta}_j$ is the component specific vector of parameters, with $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j^T, \boldsymbol{\Sigma}_j^T)$ for the multivariate normal distribution, $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j^T, \boldsymbol{\Sigma}_j^T, \boldsymbol{\lambda}_j^T)$ for the

multivariate skew-normal distribution, $\boldsymbol{\theta}_j = (\boldsymbol{\mu}_j^T, \boldsymbol{\Sigma}_j^T, \boldsymbol{\lambda}_j^T, \nu)$ for the multivariate skew-t distribution, $\boldsymbol{\mu}_j^T = (\mu_{j1},...,\mu_{jp})^T$, $\boldsymbol{\Sigma}_j^T = \begin{bmatrix} \sigma_{j,11} & ... & \sigma_{j,1p} \\ ... & ... & ... \\ \sigma_{j,p1} & ... & \sigma_{j,pp} \end{bmatrix}$, $\boldsymbol{\lambda}_j^T = (\lambda_{j1},...,\lambda_{jp})^T$, and $\psi(\mathbf{y} \mid \boldsymbol{\theta}_j) =$ multivariate normal, skew-normal or skew-t density function.

In the mixture context, we consider the latent component-indicator variables $\mathbf{Z}_i = (Z_{i1},...Z_{ig})^T$, $i = 1,...,n$, to classify each vector observation $\mathbf{y}_i$, which is defined as

$$Z_{ij} = \begin{cases} 1, \text{ if } \mathbf{y}_i \text{ belongs to group j,} \\ 0, \text{ otherwise} \end{cases} \tag{5.8}$$

and $\sum_{j=1}^{g} Z_{ij} = 1$. $\mathbf{Z}_1,...,\mathbf{Z}_n$ are independent random vectors and each of them has a multinomial distribution with density given:

$$f(\mathbf{z}_i) = w_1^{z_{i1}} w_2^{z_{i2}} ...(1 - w_1 - ... - w_{g-1})^{z_{ig}} \tag{5.9}$$

Thus, we denote it as $\mathbf{Z}_i \sim M(1; w_1,..., w_g)$.

5.4 Model Estimation Method

Compared with the normal mixture model, the parameter estimation process is more challenging for the skew-normal and skew-t mixture models. Lin et al. (2007) and Lin (2010) implemented the maximum likelihood estimation of the univariate and multivariate skew-t mixture models via a modified Expectation-Maximization (EM) algorithm. Recently, Cabral et al. (2012) also developed a general EM-type algorithm for

estimating parameters of finite mixtures of multivariate skew-normal and skew-t distributions. Since most studies on finite mixtures of multivariate distribution employed the maximum likelihood estimation with EM algorithm, we also compute the maximum likelihood estimates for the model parameters. For more details about the EM algorithm used in this chapter, interested readers can see Cabral et al. (2012).

5.5 Goodness of Fit Statistics

To evaluate the goodness of fit of the selected mixture models, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), R^2 and root mean square error (RMSE) statistics are used.

The AIC and BIC have the same form $-2LL + \gamma c_n$, where LL is the log-likelihood value, γ is the number of free parameters to be estimated and c_n is the penalty term with a positive value. The value of c_n is defined depending on the selected criterion. For the AIC and BIC, c_n equals 2 and $\log(n)$ respectively, where n is the number of observations in the data. In the density estimation context, the BIC is a reliable tool for comparing mixture models.

R^2 statistic is a bin-specific test. The common definition of the R^2 is

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \tag{5.10}$$

where SS_{err} represents the sum of squares of the residuals and SS_{tot} denotes the total sum of squares. R^2 statistic ranges from 0 to 1 and higher R^2 values indicate a better fit.

The RMSE statistic is also bin-specific and has the following form:

$$RMSE = \sqrt{\frac{SS_{err}}{N_T}} \tag{5.11}$$

where SS_{err} represents the sum of squares of residuals, and N_T is the total number of bins. Unlike the R^2 statistic, higher RMSE values indicate a poorer fit. Note that when calculating the R^2 and RMSE statistics for the bivariate distribution, SS_{err} reflects the total difference between the observed and expected frequency for all of the two-dimensional bins, and N_T is the total number of two-dimensional bins. For speed, the bin size of R^2 metric is fixed at 2 kph, whereas for headway, the bin size is specified as 1 second. The RMSE metric uses the same bin size.

5.6 Modeling Results

We apply bivariate normal, skew-normal and skew-t mixture models with an increasing number of components (g = 2,…,6) to the 24-hour speed and headway data described in Chapter III. A common problem with the EM algorithm is that the likelihood function of mixture models might have multiple roots corresponding to local maxima (Zou et al., 2012). Thus, in order to ensure a global maximum has been found, many different random starting values are applied with the EM algorithm and we select the optimal estimation result that corresponds to the largest likelihood value (Zou et al., 2012).

5.6.1 Determination of the optimal model

To select the most appropriate model for speed and headway data from bivariate normal, skew-normal and skew-t mixture models, the AIC, the BIC, R^2 and RMSE are computed for each mixture model. Table 5.1 provides the goodness of fit statistics (i.e., Log-likelihood (LL), AIC, BIC, R^2 and RMSE) for three mixture models with g = 2,…,6. Larger LL and R^2 and smaller AIC, BIC and RMSE values indicate a better overall fit. When the number of components in the finite mixture model is small (i.e., g = 2, 3), the bivariate skew-t mixture model can provide a significant better fitting result for the speed and headway data than the other two mixture models. On the other hand, as the number of components increases, the differences of the fitting performance among three mixture models become less obvious. Overall, the bivariate skew-t mixture model can consistently outperform the bivariate normal and skew-normal mixture models in terms of the LL, AIC and BIC values while the bivariate normal mixture model provides the least satisfactory fitting performance. Based on the goodness of fit statistics in Table 5.1, we select the bivariate skew-t mixture model as the optimal model for describing the speed and headway data. The parameter estimation results for the bivariate skew-t mixture models are provided in Table 5.2. Since the 24-hour traffic data used in this research consists of distinct traffic flow conditions, the correlation structure between speed and headway varies based on the traffic condition (for example, as shown in Table 3.1, speed and headway usually have an inverse relationship during the peak period and a positive relationship during the off-peak period.). The finite mixtures of bivariate skew-t distributions can address this issue naturally, since each component has its own

covariance matrix and the correlation structure between speed and headway can be different across components.

Table 5.1 Goodness of fit statistics for three mixture models

G = 2	Normal	Skew-normal	Skew-t
LL	-192922	-184068	**-174635**
AIC	385865	368157.8	**349292.4**
BIC	385955.6	368248.4	**349383**
R^2	0.493895	0.660278	**0.92225**
RMSE	26.62933	21.81735	**10.43732**
g = 3	Normal	Skew-normal	Skew-t
LL	-176937	-174820	**-170936**
AIC	353908.3	349674.7	**341906.1**
BIC	354048.3	349814.7	**342046.1**
R^2	0.853	0.837	**0.955**
RMSE	14.319	15.097	**7.887**

Table 5.1 Continued

g = 4	Normal	Skew-normal	Skew-t
LL	-173822	-171824	**-170666**
AIC	347689.8	343693.3	**341377.3**
BIC	347879.3	343882.7	**341566.8**
R^2	0.894	0.937	**0.972**
RMSE	12.175	9.388	**6.247**
g = 5	Normal	Skew-normal	Skew-t
LL	-171360	-171727	**-170089**
AIC	342778.7	343512.4	**340236.5**
BIC	343017.6	343751.3	**340475.4**
R^2	0.962	0.956	**0.962**
RMSE	7.223	7.765	**7.220**
g = 6	Normal	Skew-normal	Skew-t
LL	-171407	-170437	**-170110**
AIC	342884	340943.5	**340290.4**
BIC	343172.3	341231.8	**340578.7**
R^2	**0.967**	0.948	0.950
RMSE	**6.737**	8.501	8.344

Table 5.2 Parameter estimation results for the bivariate Skew-t mixture models

Number of components	Parameter	Component					
		1	2	3	4	5	6
g=2	μ_{j1}	98.47	15.43				
	μ_{j2}	0.85	2.94				
	$\sigma_{j,11}$	7.71	17.77				
	$\sigma_{j,12}$	-0.13	-0.47				
	$\sigma_{j,22}$	1.85	1.27				
	λ_{j1}	-0.81	1.89				
	λ_{j2}	2.02	0.87				
	ν	2.44	2.44				
	w_j	0.85	0.15				
g=3	μ_{j1}	95.86	15.40	92.77			
	μ_{j2}	0.72	2.73	0.84			
	$\sigma_{j,11}$	5.86	8.44	18.91			
	$\sigma_{j,12}$	0.41	-0.17	-0.45			
	$\sigma_{j,22}$	2.31	1.95	0.57			
	λ_{j1}	0.73	1.19	-1.89			
	λ_{j2}	2.57	1.61	1.34			
	ν	2.82	2.82	2.82			
	w_j	0.66	0.10	0.23			
g=4	μ_{j1}	99.11	15.42	68.59	98.72		
	μ_{j2}	1.06	2.68	1.04	2.09		
	$\sigma_{j,11}$	8.39	7.11	20.22	5.40		
	$\sigma_{j,12}$	-0.22	-0.24	-0.63	0.04		
	$\sigma_{j,22}$	0.91	2.22	0.70	3.86		
	λ_{j1}	-1.23	0.93	-1.32	0.78		
	λ_{j2}	1.14	1.75	0.97	2.26		
	ν	3.87	3.87	3.87	3.87		
	w_j	0.58	0.09	0.08	0.24		

Table 5.2 Continued

g=5	μ_{j1}	31.76	98.93	99.90	14.76	83.52	
	μ_{j2}	1.88	0.98	2.92	2.73	0.84	
	$\sigma_{j,11}$	11.99	6.68	5.54	6.25	13.98	
	$\sigma_{j,12}$	-0.02	-0.23	-0.29	-0.22	-0.42	
	$\sigma_{j,22}$	0.85	1.21	4.63	2.31	0.65	
	λ_{j1}	1.20	-0.93	0.76	0.78	-1.47	
	λ_{j2}	0.95	1.56	2.28	1.73	1.41	
	ν	4.01	4.01	4.01	4.01	4.01	
	w_j	0.04	0.60	0.16	0.09	0.12	
g=6	μ_{j1}	94.20	72.57	27.21	17.05	96.96	100.41
	μ_{j2}	0.87	0.94	2.00	2.20	1.43	4.55
	$\sigma_{j,11}$	7.16	9.66	12.64	5.20	5.51	6.25
	$\sigma_{j,12}$	-0.34	-0.42	-0.12	-0.85	0.13	-0.56
	$\sigma_{j,22}$	0.78	0.69	1.00	3.00	1.46	6.32
	λ_{j1}	-1.44	-1.34	1.05	-0.64	1.12	0.62
	λ_{j2}	1.12	1.04	1.12	2.60	1.32	2.29
	ν	4.37	4.37	4.37	4.37	4.37	4.37
	w_j	0.30	0.05	0.06	0.07	0.43	0.09

The number of components in finite mixtures of bivariate distributions can be determined in two approaches: the first method is to assume that g is an unknown variable and it is estimated within the modeling process; the second way is to fit a series of models with increasing numbers of components and we select the most plausible model by the model choice criteria (Park et al., 2010). For finite mixtures of univariate distributions, some methodologies (for example, reversible jump Markov Chain Monte

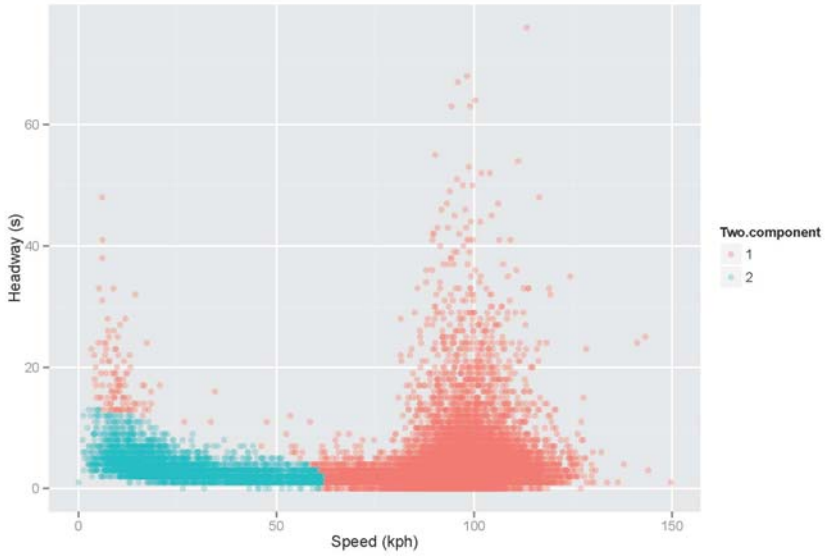
Carlo) have been proposed for the analysis of mixture models with unknown number of components. However, for finite mixtures of multivariate distributions, the implementation of the first method turns out to be very complicated and some issues remain unsolved. Thus, we adopted the model choice criteria. In this section, the bivariate skew-t distribution is selected as the component density for determining the number of components in the mixture model.

To select the optimal number of components, the information-based criteria (AIC and BIC) and classification results from the modeling process are considered. As shown in Table 5.1, the AIC and BIC values of the model with g = 2 are significantly larger than other models, indicating the assumption of two components cannot adequately capture the heterogeneity of this dataset. Thus, based on the information-based criteria, the model with g = 2 can be excluded from further consideration. Classification or grouping results were used to examine if the finite mixture model can reasonably separate the speed and headway data into different clusters. Each speed and headway data pair was classified into different groups by assigning each observation to the component with the highest posterior probability (Park et al. 2010). The posterior probability is used to calculate the probability that observation $\mathbf{y}_i$ is from component j. In the EM algorithm, at iteration r+1, the posterior probability $\hat{\varepsilon}_{ij}^{(r+1)}$ that observation $\mathbf{y}_i$ is from component j, given $\mathbf{y}_i$ and $\hat{\boldsymbol{\Theta}}^{(r)}$ is defined as (Cabral et al., 2012):

$$\hat{\varepsilon}_{ij}^{(r+1)} = p(Z_{ij}=1 \mid \mathbf{y}_i, \hat{\boldsymbol{\Theta}}^{(r)}) = \frac{\hat{w}_j^{(r)} \, f_j(\mathbf{y}_i \mid \hat{\boldsymbol{\theta}}_j^{(r)})}{\sum_{k=1}^{g} \hat{w}_k^{(r)} \, f_k(\mathbf{y}_i \mid \hat{\boldsymbol{\theta}}_k^{(r)})} \tag{5.12}$$

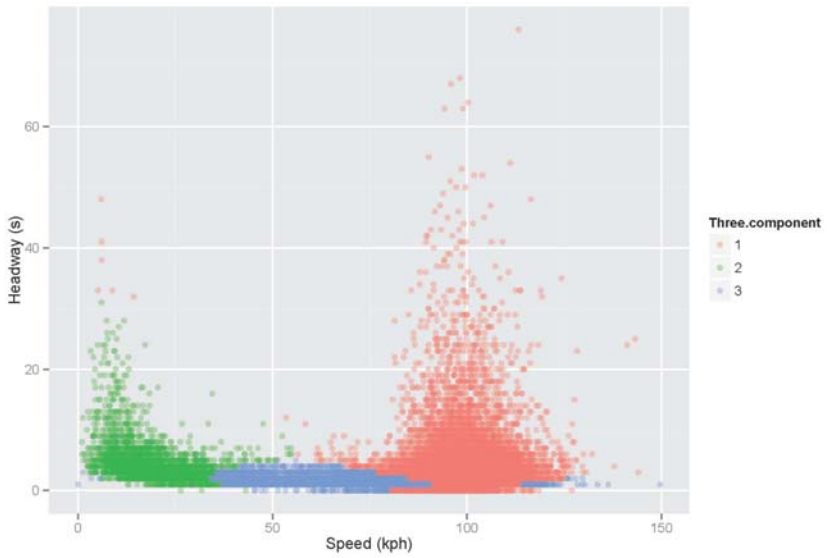
where Z_{ij} is the indicator variable, $f_j(\mathbf{y}_i \mid \hat{\boldsymbol{\theta}}_j^{(r)})$ is the component density, and $\hat{w}_j^{(r)} = p(Z_{ij}=1 \mid \hat{\boldsymbol{\Theta}}^{(r)})$ is the prior probability that observation $\mathbf{y}_i$ is from component j, given $\hat{\boldsymbol{\Theta}}^{(r)}$, which is estimated from iteration r.

Figure 5.1 shows the classification results from bivariate skew-t mixture models with g = 2,…,6. As discussed in the above paragraph, the model with g = 2 is not a viable option due to its unsatisfactory fitting performance. When we compare Figure 5.1 (b) and (c), it can be observed that component 1 in Figure 5.1 (b) is approximately and unnecessarily further separated into two sub-clusters (components 1 and 4 in Figure 5.1 (c)). Similarly, component 3 in Figure 5.1 (c) contains two sub-clusters (components 1 and 5 in Figure 5.1 (d)) and component 5 in Figure 5.1 (d) roughly consists of two sub-clusters (components 1 and 2 in Figure 5.1 (e)). Thus, for the principle of model parsimony, the three-component bivariate skew-t mixture model is preferred. For Figure 5.1 (b), the first component (red dots) represents mostly the free flow traffic condition and the second component (green dots) represents mostly the traffic condition during the peak periods. The third component (blue dots) can be viewed as the transition flow condition. The classification results shown in Figure 5.1 indicate that the heterogeneity for the 24-hour speed and headway data mainly resulted from different traffic conditions.

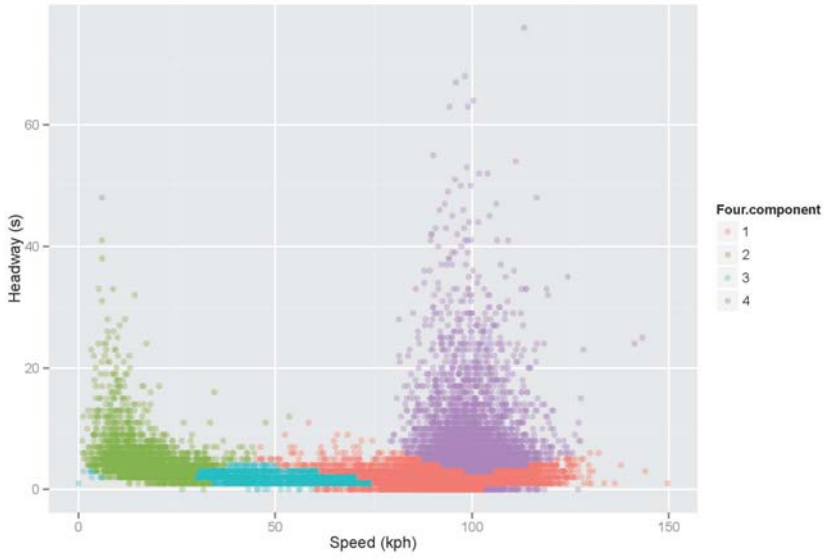


(a)

Figure 5.1 Scatter plots of grouping results from the (a) two-component; (b) three-component; (c) four-component; (d) five-component and (e) six-component bivariate skew-t mixture models.

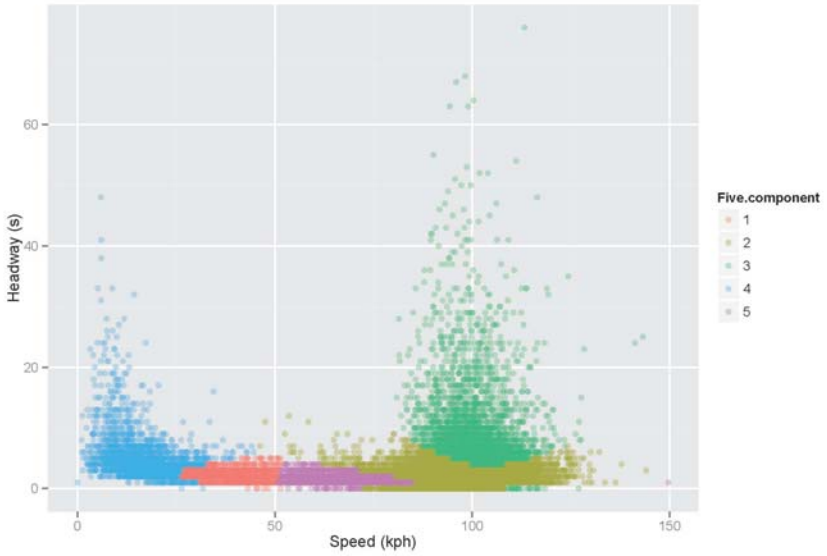


(b)



(c)

Figure 5.1 Continued



(d)

(e)

Figure 5.1 Continued

5.6.2 Effect of vehicle type on following headway

To investigate the impact of the vehicle composition on component grouping, the speed and headway data were classified into different groups by assigning each observation to the component with the highest posterior probability (Park et al., 2010). In this chapter, a long vehicle (LV) is defined as a vehicle with its length larger than or equal to 12.2 meters (40 feet). The percentages of vehicle composition of the dataset used in this paper are 10.4% (long vehicles) and 89.6% (other types of vehicles). If there exists an association between component grouping and vehicle classification, then it is expected that the proportion of LVs will be different between components. However, as indicated in Figure 5.1, it is clear that the different flow conditions are the main cause for the component separation for the 24-hour dataset. To minimize the effect of the traffic condition on speed and headway, one reasonable method to investigate the influence of vehicle type on speed and headway is to confine the analysis to traffic data with a specified speed range. For the congested traffic condition, Ye and Zhang (2009) and Sarvi (2011) showed that passenger cars take a longer time headway behind long vehicles than when following other passenger cars. Similarly, long vehicles also adopt longer headway (in time) when following other vehicles due to their less agile operating characteristics with respect to acceleration and deceleration. Since the influence of vehicle type on speed and headway is more obvious in the car following situation, further analysis was carried out using the traffic data observed in the congested traffic condition. Specifically, we consider the traffic data from five different speed groups (i.e., 0-10 kph, 10-20 kph, …, 40-50 kph). For each sub-dataset within the specified speed

range, the two-component bivariate skew-t mixture model was applied and the classification results were provided in Table 5.3. By assuming two components, it is helpful to understand what factors make a particular observation more prone to fall into one or the other sub-population (Zou et al., 2012).

For speed group 1 with speed values less than 10 kph, there is a difference in the average value of headway between the two components; while there is no significant difference in the average value of speed between two components. Table 5.3 shows that the following or leading vehicles (especially the leading vehicles) in component 1 are more likely to be long vehicles than component 2. The classification results for speed group 1 suggest that larger following or leading vehicle length generally results in longer time headway. Similar findings can be observed for group 2. Interestingly, for groups with speed values greater than 30 kph, the effect of vehicle type on headway and speed is not as significant. The scatter plot of speed and headway illustrated in Figure 3.2 can be seen as evidence to support the findings drawn from the mixture modeling. As shown in Figure 3.2, the speed and headway data points are highly dispersed when speed values are below 20 kph and gradually become concentrated as speed increases. Thus, there should be some factors (for example, vehicle type) to explain this interesting pattern. Overall, the analysis in this part shows that the bivariate skew-t mixture modeling approach has the flexibility in explaining the impact of other factors (for example, vehicle type) on speed and headway.

Table 5.3 Effect of vehicle type on headway and speed under the congested traffic condition

Speed groups (kph)	Component	Average speed	Average headway	Average vehicle length		Percentage of LVs	
				following	leading	following	leading
1 (0-10 kph)	1 (73)[a]	7.535	18.52	7.843	12.12	19.2%	42.5%
	2 (375)	7.209	5.568	6.586	6.021	11.5%	9.1%
2 (10-20 kph)	1 (1152)	15.12	3.761	6.454	5.629	11.5%	6.5%
	2 (145)	14.28	11.34	8.91	14.11	26.9%	57.9%
3 (20-30 kph)	1 (569)	22.46	3.42	6.349	6.506	11.2%	12.3%
	2 (349)	27.59	2.739	5.794	5.787	7.7%	7.2%
4 (30-40 kph)	1 (290)	32.34	2.869	6.478	6.429	11.7%	11.7%
	2 (202)	37.55	2.322	5.669	5.623	7.4%	6.9%
5 (40-50 kph)	1 (296)	42.87	2.135	5.634	5.904	7.1%	9.1%
	2 (162)	47.59	2.352	6.083	5.776	10.5%	7.4%

Note: [a] Number of observations in each component.

5.7 Summary

Although finite mixtures of univariate distributions can capture the heterogeneity observed in one-dimensional data (i.e., speed data), this modeling approach neglects the possible correlation between speed and headway. This chapter examined the

applicability of the finite mixtures of multivariate distributions to accommodate the heterogeneity existing in speed and headway data. It is found that the bivariate skew-t mixture model can provide a satisfactory fit to the speed and headway distribution and this modeling approach can accommodate the varying correlation coefficient. For the 24-hour freeway speed and headway data, the three-component bivariate skew-t mixture model was considered as the optimal model. For the speed and headway data observed under the congested traffic condition, the use of the bivariate skew-t mixture model demonstrated that vehicle type has a significant impact on following headway when speed is below 20 kph.

CHAPTER VI

METHODOLOGY III: MODELING FREEWAY SPEED AND HEADWAY USING COPULAS

6.1 Introduction

In the previous chapter, the bivariate skew-t mixture model was proposed to describe the speed and headway data. Although bivariate skew-t distribution can accommodate dependence structure between speed and headway, the main restriction of this approach is that the individual behavior of speed and headway is characterized by the same univariate distributions. Therefore, this chapter introduces copula models which can avoid this restriction.

6.2 Concept of Copulas

The concept of copula was first proposed by Sklar (1959) and the interests in copulas and their application in the statistics field have grown over the last decades (see Genest and MacKay (1986); Genest and Rivest (1993); Nelsen (2006)). Recently, the copula method has received much attention from the finance, hydrological modeling, econometrics and transportation fields (see, Embrechts et al. (2002); Cherubini et al. (2004); Zhang and Singh (2006); Bhat and Eluru (2009)).

What are copulas? Copulas are functions that join or "couple" multivariate distribution functions to their one-dimensional marginal distribution functions (Nelsen, 2006). For

continuous random variables X and Y, the Sklar's theorem (1959) stated that let $H(x,y)$ be a joint cumulative distribution function (cdf) with continuous marginal distributions $F(x)$ and $G(y)$, then there exists a bivariate copula C:

$$H(x,y) = C(F(x), G(y)) \tag{6.1}$$

where $C:[0,1]^2 \to [0,1]$ = copula.

A valid model for (X,Y) can be obtained from equation (6.1) if $F(x)$ and $G(y)$ are selected from parametric families of distributions. For example, $F(x)$ can be a normal distribution with parameters (μ,σ^2) and $G(y)$ can be an exponential distribution with parameter λ. Moreover, a rich set of copula types C are available for generating the joint cdf $H(x,y)$. These copula types include the Gaussian copula, the Farlie-Gumbel-Morgenstern copula, and various Archimedean copulas (a detailed introduction to these copulas is provided in section 6.4). One advantage of the copula approach is that the selection of a model for representing X and Y can proceed independently from the choice of the marginal distributions (Genest and Favre, 2007).

For continuous distribution functions $F(x)$ and $G(y)$, the generalized inverse functions are defined by $F^{-}(t) = \inf\{x \mid F(x) \geq t\}$ and $G^{-}(t) = \inf\{y \mid G(x) \geq t\}$, respectively. Let $U = F(X)$ and $V = G(Y)$, then based on the probability integral transform, U and V are uniformly distributed random variables with support $[0, 1]$. We can obtain

$$F(x) = \Pr(X < x) = \Pr\left(F^{-1}(U) < x\right) = \Pr\left(U < F(x)\right) \quad \text{and}$$

$$G(y) = \Pr(Y < y) = \Pr\left(G^{-1}(V) < y\right) = \Pr\left(V < G(y)\right).$$

Let $H(x, y)$ be a distribution function with continuous marginal distributions $F(x)$ and $G(y)$, then for any $u, v \in [0, 1]$, the copula function can be defined as (Nelsen, 2006):

$$C(u, v) = H\left(F^{-}(u), G^{-}(v)\right) \tag{6.2}$$

6.3 Measuring Dependence

There are different ways to measure dependence. Some measures are scale-invariant (i.e., these measures remain unchanged under strictly increasing transformations of the random variables). Two widely known scale-invariant measures of association are Kendall's tau and Spearman's rho. Specifically, let (x_i, y_i) and (x_j, y_j) be two observations from a vector (X, Y) of continuous random variables. It is defined that (x_i, y_i) and (x_j, y_j) are concordant if $x_i < y_i$ and $x_j < y_j$, or if $x_i > y_i$ and $x_j > y_j$. Similarly, (x_i, y_i) and (x_j, y_j) are discordant if $x_i < y_i$ and $x_j > y_j$, or if $x_i > y_i$ and $x_j < y_j$.

Assume (X_1, Y_1) and (X_2, Y_2) be independent and identically distributed random vectors, with a joint distribution function $H(x, y)$. The population version of Kendall's

tau can be defined as the probability of concordance minus the probability of discordance (Nelsen, 2006):

$$\tau_{X,Y} = P[(X_1 - X_2)(Y_1 - Y_2) > 0] - P[(X_1 - X_2)(Y_1 - Y_2) < 0] \tag{6.3}$$

Let X and Y be continuous random variables whose copula is C. For $H(x,y) = C(u = F(x), v = G(y))$, the expression of Kendall's tau $\tau_{X,Y}$ above can be rewritten as (see Nelsen, 2006, p. 159-162 for a proof):

$$\tau_{X,Y} = 4\iint_{[0,1]^2} C(u,v)\, dC(u,v) - 1 \tag{6.4}$$

Let (X_1, Y_1), (X_2, Y_2), and (X_3, Y_3) be three independent random vectors with a common joint distribution function $H(x,y) = C(u = F(x), v = G(y))$. The population version of Spearman's rho $\rho_{X,Y}$ is proportional to the probability of concordance minus the probability of discordance for the two vectors (X_1, Y_1) and (X_2, Y_3), which is given by:

$$\rho_{X,Y} = 3(P[(X_1 - X_2)(Y_1 - Y_3) > 0] - P[(X_1 - X_2)(Y_1 - Y_3) < 0]) \tag{6.5}$$

Note that the joint distribution function of (X_1, Y_1) is $H(x,y)$, and X_2 and Y_3 are independent, indicating the joint distribution function of (X_2, Y_3) is $F(x)G(y)$.

Let X and Y be continuous random variables whose copula is C. For $H(x,y) = C(u = F(x), v = G(y))$, the expression of Spearman's rho $\rho_{X,Y}$ above can be rewritten as (see Nelsen, 2006, p. 167 for a proof):

$$\rho_{X,Y} = 12\iint_{[0,1]^2} uvdC(u,v) - 3 = 12\iint_{[0,1]^2} C(u,v)dudv - 3 \tag{6.6}$$

Besides the Kendall's tau and Spearman's rho, one traditional correlation coefficient needs to be mentioned is the Pearson's product-moment correlation coefficient, which measures the linear dependence between random variables. Compared with the rank-based correlation, the linear correlation has the deficiency that it is not invariant under nonlinear strictly increasing transformations (Embrechts et al., 2002). Embrechts et al. (2002) also pointed out that for multivariate distributions which possess a simple closed-form copula, the moment-based correlations (i.e. Pearson's correlation coefficient) may be difficult to calculate and the determination of rank-based correlation (i.e., Kendall's tau and Spearman's rho) may be easier. Therefore, considering the advantages of rank-based correlation, the Kendall's tau and Spearman's rho are used to characterize the dependence structure for different types of copulas described in the following section.

6.4 Family of Bivariate Copulas

6.4.1 Bivariate Gaussian copulas

The Gaussian copula can be obtained using the inversion method. The 2-dimensional Gaussian copula with linear correlation matrix Σ is given by:

$$C_{\Sigma}(u,v) = \Phi_{\Sigma}\left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$$
$$= \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(-\frac{s^2 - 2\theta st + t^2}{2\left(1-\theta^2\right)}\right) dsdt \tag{6.7}$$

where $\Sigma = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}$ is the correlation matrix, with parameter $\theta \in (-1, 1)$, Φ_{Σ} is a standard bivariate normal distribution and Φ is a standard normal distribution. If $\theta = 0$, the Gaussian copula becomes to the independent copula. Dependence parameter θ and Kendall's tau have the relationship, that is, $\tau = (2/\pi)\sin^{-1}(\theta)$. The 2-dimensional Gaussian copula density function is given by:

$$c_{\Sigma}(u,v) = \frac{1}{\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\omega^T\left(\Sigma^{-1} - I_2\right)\omega\right) \tag{6.8}$$

where $\omega^T = \left(\Phi^{-1}(u), \Phi^{-1}(v)\right)$, I_2 is the 2×2 identity matrix.

6.4.2 The Farlie-Gumbel-Morgenstern copula

The FGM was originally introduced by Morgenstern for Cauchy marginals and investigated by Gumbel for exponential marginals, and later generalized to arbitrary functions by Farlie. The FGM copula is an intuitive and natural way to construct the joint distribution function based on the marginal cdf. The joint cdf of a bivariate distribution constructed by the FGM copula can be described as follows:

$$C_{\theta}(u,v) = uv[1 + \theta(1-u)(1-v)] \tag{6.9}$$

where θ is a parameter of the copula function and for absolutely continuous marginal distributions, we need $|\theta| \leq 1$ (Schucany et al., 1978).

And the density of the FGM copula is provided by:

$$c_\theta(u,v) = [1+\theta(2u-1)(2v-1)] \tag{6.10}$$

The FGM copula has the limitation that only if the correlation of two variables is weak, the FGM can provide an effective way for constructing a bivariate distribution. The correlation structure of FGM copula has been investigated for various continuous marginal distributions such as uniform, normal, exponential, gamma and Laplace distributions. For the rank-based dependence measures, Schucany et al. (1978) showed that, regardless of the forms of marginal distributions, θ and concordance-based correlation ($\tau_{X,Y}$ and $\rho_{X,Y}$) satisfy the following equations:

$$\tau_{X,Y} = \frac{2}{9}\theta \tag{6.11}$$

$$\rho_{X,Y} = \frac{\theta}{3} \tag{6.12}$$

Since θ is in $[-1,1]$, the FGM copula can allow weak positive and negative dependence and $\tau_{X,Y}$ and $\rho_{X,Y}$ are bounded on $[-\frac{2}{9},\frac{2}{9}]$ and $[-\frac{1}{3},\frac{1}{3}]$, respectively.

6.4.3 Bivariate Archimedean copulas

Archimedean copulas are important class of copulas and these copulas are widely applied for a few reasons: (1) Archimedean copulas have a simple and explicit form

expression; (2) they are characterized by a single parameter function φ that meets certain requirements; (3) a variety of families of copulas which belong to this class. Archimedean copulas were introduced by Genest and MacKay (1986). One parameter Archimedean copulas are briefly introduced in the following paragraph, further details can be found in Nelsen (2006).

As defined in Nelsen (2006), let φ be a continuous, strictly decreasing function from $[0,1]$ to $[0,\infty]$ such that $\varphi(1)=0$. The pseudo-inverse of φ is the function $\varphi^{[-1]}:[0,\infty]\to[0,1]$ such that $\varphi^{[-1]}(t)=\begin{cases}\varphi^{-1}(t) & 0\le t\le\varphi(0)\\ 0 & \varphi(0)\le t\le\infty\end{cases}$. If we assume $\varphi(0)=\infty$, then $\varphi^{[-1]}=\varphi^{-1}$, and we have $\varphi(\varphi^{[-1]}(t))=t$. Using functions φ and φ^{-1}, the definition of one parameter Archimedean copulas is given as:

$$C_\theta(u,v)=\varphi^{-1}(\varphi(u)+\varphi(v)) \tag{6.13}$$

The function φ is called a generator of the copula. When $\varphi(0)=\infty$, φ is said to be a strict generator and $C_\theta(u,v)$ in Equation (6.13) is a strict Archimedean copula. In the following paragraphs, several well-known one-parameter families of Archimedean copulas, along with their generators are described.

6.4.3.1 Ali-Mikhail-Haq copula

The Ali-Mikhail-Haq copula, proposed by Ali et al. (1978), can allow for weak positive and negative dependence. The generator function is $\varphi(t) = \ln \frac{1-\theta(1-t)}{t}$, with $\theta \in [-1,1)$, and the corresponding Ali-Mikhail-Haq copula function is as follows:

$$C_\theta(u,v) = \frac{uv}{1-\theta(1-u)(1-v)} \tag{6.14}$$

Kendall's tau is related to θ by $\tau = \frac{3\theta-2}{3\theta} - \frac{2(1-\theta)^2}{3\theta^2}\ln(1-\theta)$, so that $-0.182 < \tau < 0.333$. The density function of Ali-Mikhail-Haq copula is given by (Hofert et al., 2012):

$$c_\theta(u,v) = \frac{(1-\theta)^3}{\theta^2}\frac{h_\theta^A(u,v)}{u^2v^2} Li_{-2}\left\{h_\theta^A(u,v)\right\} \tag{6.15}$$

where $h_\theta^A(u,v) = \theta \frac{u}{1-\theta(1-u)} \frac{v}{1-\theta(1-v)}$ and $Li_s(z) = \sum_{k=1}^{\infty} z^k / k^s$.

6.4.3.2 The Clayton copula

If the generator function is selected as $\varphi(t) = \frac{1}{\theta}(t^{-\theta}-1)$, with $\theta \in (0,\infty)$, the Archimedean copula is called the Clayton copula. It is given by:

$$C_\theta(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \tag{6.16}$$

The Clayton copula was first proposed by Clayton (1978) and allows only positive dependence. Kendall's tau is related to θ by $\tau = \frac{\theta}{\theta+2}$, so that $0 < \tau < 1$. If θ tends to 0, the Clayton copula becomes independent copula. The density function of Clayton copula is given by (Hofert et al., 2012):

$$c_\theta(u,v) = (1+\theta)u^{-(\theta+1)}v^{-(\theta+1)}(u^{-\theta}+v^{-\theta}-1)^{-1/\theta-2} \tag{6.17}$$

6.4.3.3 The Frank copula

If we choose $\varphi(t) = -\ln\frac{e^{-\theta t}-1}{e^{-\theta}-1}$, with $\theta \in (-\infty,\infty)\setminus\{0\}$, the Archimedean copula is called the Frank copula. It is given by:

$$C_\theta(u,v) = -\frac{1}{\theta}\ln\left(1+\frac{\left(e^{-\theta u}-1\right)\left(e^{-\theta v}-1\right)}{e^{-\theta}-1}\right) \tag{6.18}$$

with $\tau = 1-\frac{4}{\theta}[1-D_1(\theta)]$, where $D_1(\theta)$ is the first order Debye function $D_k(\theta)$ which is defined as $D_k(\theta) = \frac{k}{\theta^k}\int_0^\theta \frac{t^k}{e^t-1}dt$.

The Frank copula was proposed by Frank (1979). It can allow for both positive and negative dependence. The range of τ is (−1, 1) and if θ tends to 0, the Frank copula becomes independent copula. The density function of Frank copula is given by (Hofert et al., 2012):

$$c_\theta(u,v)=\left(\frac{\theta}{1-e^{-\theta}}\right)Li_{-1}\left\{h_\theta^F\left(u,v\right)\right\}\frac{\exp(-\theta(u+v))}{h_\theta^F\left(u,v\right)} \tag{6.19}$$

where $h_\theta^F\left(u,v\right)=(1-e^{-\theta})^{-1}(1-\exp(-\theta u))(1-\exp(-\theta v))$.

6.4.3.4 The Gumbel copula

The Gumbel copula, also known as the Gumbel-Hougaard copula, was first introduced by Gumbel (1960). The generator function for this copula is $\varphi(t)=(-\ln t)^\theta$, and the corresponding copula function is

$$C_\theta(u,v)=\exp\left(-\left[\left(-\ln u\right)^\theta+\left(-\ln v\right)^\theta\right]^{1/\theta}\right) \tag{6.20}$$

The Gumbel copula only accommodates positive dependence and Kendall's tau is related to θ by $\tau=1-\theta^{-1}$, so that $0<\tau<1$. If $\theta=1$, the Gumbel copula becomes independent copula. The density function of Gumbel copula is given by (Hofert et al., 2012):

$$c_\theta(u,v)=\theta^2\exp\left\{-t_\theta\left(u,v\right)^\alpha\right\}\frac{(-\ln u)^{\theta-1}(-\ln v)^{\theta-1}}{t_\theta\left(u,v\right)^2uv}P_{2,\alpha}^G\left(t_\theta\left(u,v\right)^\alpha\right) \tag{6.21}$$

where $\alpha=1/\theta$, $P_{2,\alpha}^G\left(x\right)=\sum_{k=1}^{2}\varepsilon_{2k}^G(\alpha)x^k$, and $\varepsilon_{2k}^G(\alpha)=\frac{2}{k!}\sum_{j=1}^{k}\binom{k}{j}\binom{\alpha j}{2}(-1)^{2-j}$.

6.4.3.5 The Joe copula

The Joe copula, discussed by Joe (1993, 1997), has a generator function $\varphi(t) = -\ln\left[1-(1-t)^{\theta}\right]$. The Joe copula is defined as:

$$C_{\theta}(u,v) = 1-\left[(1-u)^{\theta}+(1-v)^{\theta}-(1-u)^{\theta}(1-v)^{\theta}\right]^{1/\theta} \tag{6.22}$$

with $\tau = 1+\frac{4}{\theta}D_J(\theta)$, where $D_J(\theta) = \int_{t=0}^{1}\frac{[\ln(1-t^{\theta})](1-t^{\theta})}{t^{\theta-1}}dt$

Like the Clayton and Gumbel copulas, the Joe copula can not account for negative dependence. The range of τ is (0, 1) . If θ tends to 0, the Joe copula becomes independent copula. The density function of Joe copula is given by (Hofert et al., 2012):

$$c_{\theta}(u,v) = \theta\frac{(1-u)^{\theta-1}(1-v)^{\theta-1}}{\left\{1-h_{\theta}^{J}(u,v)\right\}^{1-\alpha}}P_{2,\alpha}^{J}\left\{\frac{h_{\theta}^{J}(u,v)}{1-h_{\theta}^{J}(u,v)}\right\} \tag{6.23}$$

where $\alpha = 1/\theta$, $h_{\theta}^{J}(u,v) = \left\{1-(1-u)^{\theta}\right\}\left\{1-(1-v)^{\theta}\right\}$, $P_{2,\alpha}^{J}(x) = \sum_{k=0}^{1}\varepsilon_{2k}^{J}(\alpha)x^{k}$,

$\varepsilon_{2k}^{J}(\alpha) = S(2,k+1)\frac{\Gamma(k+1-\alpha)}{\Gamma(1-\alpha)}$ and $S(j,k)$ is the Stirling numbers of the second kind.

6.5 Multivariate Gaussian Copulas

The copula of the n-variate normal distribution with $n \times n$ correlation matrix P is

$$C_{\mathrm{P}}(\mathbf{u}) = \Phi_{\mathrm{P}}\left(\Phi^{-1}(u_1),\Phi^{-1}(u_2),...,\Phi^{-1}(u_n)\right) \tag{6.24}$$

where Φ_{P} represents the joint distribution function of the n-variate standard normal distribution function with correlation matrix P, and Φ^{-1} is the inverse of the distribution function of the univariate standard normal distribution. For the multivariate Gaussian copula, correlation matrix P and Kendall's tau have the relationship, that is $\tau_{X_i,X_j} = \frac{2}{\pi}\sin^{-1}\left(\rho_{ij}\right)$ (Embrechts et al, 2003; Demarta and McNeil , 2005).

In the trivariate case the copula expression can be written as

$$C_{\mathrm{P}}(u_1,u_2,u_3)= \int_{-\infty}^{\Phi^{-1}(u_1)}\int_{-\infty}^{\Phi^{-1}(u_2)}\int_{-\infty}^{\Phi^{-1}(u_3)} \frac{1}{(2\pi)^{3/2}\left|\mathrm{P}\right|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{w}^T\mathrm{P}^{-1}\mathbf{w}\right)d\mathbf{w} \tag{6.25}$$

where $\mathrm{P}=\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}$ is the symmetrical correlation matrix with $-1 \le \rho_{ij} \le 1$ ($i, j = 1,2,3$); $\mathbf{w} = \left(w_1, w_2, w_3\right)^T$ represents the corresponding integral variables.

Besides the multivariate Gaussian copula, multivariate Archimedean copulas are also widely used for modeling multivariate distribution of multiple random variables. The multivariate Archimedean copulas include the symmetric Archimedean copula and the asymmetric Archimedean copula (which is also called nested Archimedean copula). Note that the symmetric Archimedean copula is a special case of the asymmetric Archimedean copula. The symmetric Archimedean copula suffers from a very limited dependence structure since all k-margins are identical; they are distribution functions of

n exchangeable $U(0,1)$ random variables (Embrechts et al., 2003). As a consequence of this exchangeability property, all mutual dependences among variables are modeled by only one Archimedean 2-copula (Grimaldi and Serinaldi, 2006). On the other hand, the asymmetric Archimedean copula allows for nonexchangeability and a part of all possible mutual dependences can be modeled in a different way. For more details about multivariate Archimedean copulas, interested readers can see Grimaldi and Serinaldi (2006). Compared with the multivariate Gaussian copulas which are able to model all range of dependence, the multivariate Archimedean copula families ($n \geq 3$) can model only positive dependence. Thus, considering the possible inverse relationship between speed and headway, only multivariate Gaussian copulas are considered.

6.6 Estimation of θ

Given a parametric family (C_θ) of copulas and a random sample (X_1, Y_1), ..., (X_n, Y_n) from continuous random variables (X, Y), the first step is to select appropriate marginal distributions for each variable. Then the data can be transformed onto the copula scale using the probability integral transform. The next step is to estimate θ. Genest and Favre (2007) reviewed various nonparametric methods for estimating θ and they recommend using ranked-based estimators since the ranks of the observations are the best summary of the joint behavior of the random pairs. Two straightforward estimators are based on Kendall's Tau and Spearman's Rho. These two rank-based estimators are explained in the following example.

If the dependence structure of a random pair (X,Y) can be appropriately modeled by the FGM copula described in Equation (6.9). Thus, as discussed above, there exist relations between the parameter θ and Kendall's Tau and Spearman's Rho, which are

$$\tau_{X,Y} = \frac{2}{9}\theta \tag{6.26}$$

$$\rho_{X,Y} = \frac{\theta}{3} \tag{6.27}$$

Since $\tau_{X,Y}$ and $\rho_{X,Y}$ can be computed from the sample pairs, a simple and intuitive approach to estimating θ would be

$$\hat{\theta} = \frac{9}{2}\tau_{X,Y} \tag{6.28}$$

$$\hat{\theta} = 3\rho_{X,Y} \tag{6.29}$$

$\tau_{X,Y}$ and $\rho_{X,Y}$ are rank-based, and this estimation strategy may be seen as a nonparametric adaptation of the method of moments (Genest and Favre, 2007).

Another method for estimating θ is called the method of maximum pseudolikelihood, which requires that C_θ be absolutely continuous with density c_θ. The concept is to maximize a rank-based log-likelihood function, which takes the form:

$$\ell(\theta) = \sum_{i=1}^{n} \log\left\{ c_\theta\left(\frac{R_i}{n+1}, \frac{S_i}{n+1} \right) \right\} \tag{6.30}$$

where R_i stands for the rank of X_i among $X_1,...,X_n$, and S_i stands for the rank of Y_i among $Y_1,...,Y_n$.

Compared with Kendall's Tau and Spearman's Rho, the maximum pseudolikelihood estimator has the advantage that it does not require the dependence parameter θ to be real. However, this method also involves a lot of numerical work and requires the existence of a density c_θ. Thus, for simplicity, the Kendall's Tau based estimator is adopted in this research. For detailed procedure of using the maximum pseudolikelihood estimator, see Genest et al. (1995). Note that Joe (1997, Chap. 10) also introduced a parametric two-step procedure referred to the inference from margins (IFM) method for estimating θ. Kim et al. (2007) pointed out that the IFM estimator depends on the choice of margins, and may run the risk of being unduly affected if selection of the margins turn out to be inappropriate.

6.7 Random Variate Generation

One of the primary applications of copulas is in simulation and Monte Carlo studies (Nelson, 2006). Based on Sklar's theorem, the copula can be used as a tool for generating observations (x, y) of a pair of random variables (X, Y) from copula function C_θ with marginal distributions $F(x)$ and $G(y)$. Specifically, we need to generate uniform random variates (u, v) from the desired copula C_θ, and then use the

inverse distribution function method to transform the data, $(x, y) = \left(F^{-1}(u), G^{-1}(v)\right)$. This section describes three algorithms for copula simulation.

6.7.1 Conditional distribution method

One general procedure for generating (u, v) from a certain copula is the conditional distribution method. Before introducing the algorithm, we first define the conditional distribution function for V given $U = u$, which is given by

$$C_u(v) = \Pr(V \le v \mid U = u) = \lim_{\Delta u \to 0} \frac{C(u + \Delta u, v) - C(u, v)}{\Delta u} = \frac{\partial C(u, v)}{\partial u} \tag{6.31}$$

Then, the algorithm for generating the uniform random variates (u, v) from the copula C_θ is defined as (Nelson, 2006, p. 41):

1. Generate two independent uniform $(0,1)$ variates u and t;
2. Set $v = C_u^{-1}(t)$, where $C_u^{-1}(\bullet)$ is a generalized inverse of C_u.

6.7.2 Sampling algorithm for Gaussian copulas

For the conditional distribution method, it is necessary to obtain the partial derivative of C_θ. However, for some copulas (i.e., Gaussian copula), it is difficult to get the analytical partial derivative. Thus, a widely used algorithm for sampling from Gaussian copula is as follows:

1. Generate $(y_1, y_2)^T$ from a bivariate normal distribution $N(0, \Sigma)$, where Σ is a

correlation matrix.

2. Set $u = \Phi(y_1)$, $v = \Phi(y_2)$.

6.7.3 Sampling algorithm for Archimedean copulas

Here, we describe another procedure for sampling from Archimedean copulas. Let joint distribution function $H(s,t)$ of the random variables $S = \varphi(U)/[\varphi(U)+\varphi(V)]$ and $T = \varphi^{-1}(\varphi(U)+\varphi(V))$ is given by $H(s,t) = sK_C(t)$ for all $(s,t) \in [0,1]^2$, where $K_C(t) = t - \varphi(t)/\varphi'(t^+)$, $\varphi'(t^+)$ denotes the one-sided derivatives of φ at t (Nelson, 2006). Hence, *S* and *T* are independent, and *S* is uniformly distributed on $[0,1]$ (for a proof, see Nelson, 2006, p. 129). Then the algorithm for generating random variates (u,v) is given by:

1. Generate two independent uniform $(0,1)$ variates s and t;

2. Set $w = K_C^{(-1)}(t)$;

3. Set $u = \varphi^{-1}(s\varphi(w))$, $v = \varphi^{-1}((1-s)\varphi(w))$.

6.8 Dependence between Microscopic Traffic Variables

Vehicle type is known as an important factor in the car following situation. For example, some studies (Ye and Zhang, 2009; Sarvi, 2011) showed that passenger cars usually travel further behind long vehicles than when following short vehicles and long vehicles also take longer time headways when following other vehicles due to their less agile

operating characteristics. In this section, to consider dependence structure among speed, headway and vehicle length, we construct a multivariate distribution of these three traffic variables. We first examine their dependence structure among each other. Since the 24-hour traffic data collected on IH-35 consists of distinct traffic flow conditions, it is possible that the dependence structure between traffic variables may vary depending on the traffic condition. Thus, we first evaluate the hourly dependence among speed, headway and vehicle length for the 24-hour period. For each hour, Kendall's tau τ , and Spearman's rho ρ_S are used to measure the dependence. The computed values of Kendall's tau τ , and Spearman's rho ρ_S for each of the 24-hour are given in Table 6.1.

As shown in Table 6.1 below, the dependence structure among three traffic variables exhibits different characteristics. First, for speed and headway, the dependence structure is stable under the same traffic condition, but change significantly between different traffic conditions. Generally speaking, for the off-peak period, when the flow rate is below 1000 vehicles/hour (i.e., 00:00 to 06:00 and 23:00 to 24:00), Kendall's tau values indicate that speed and headway have negligible effect on each other; when the flow rate is above 1000 vehicles/hour (i.e., 06:00 to 07:00, 09:00 to 15:00 and 20:00 to 23:00), τ ranges between 0.08 and 0.15 and speed and headway have a very weak positive correlation. On the other hand, for the peak period, when the flow rate is below 1000 vehicles/hour (i.e., 16:00 to 19:00), speed and headway have a weak negative dependence. Note that compared to the afternoon peak period (most speed values are below 40 kph from 16:00 to 19:00), the morning peak period (a large portion of speed

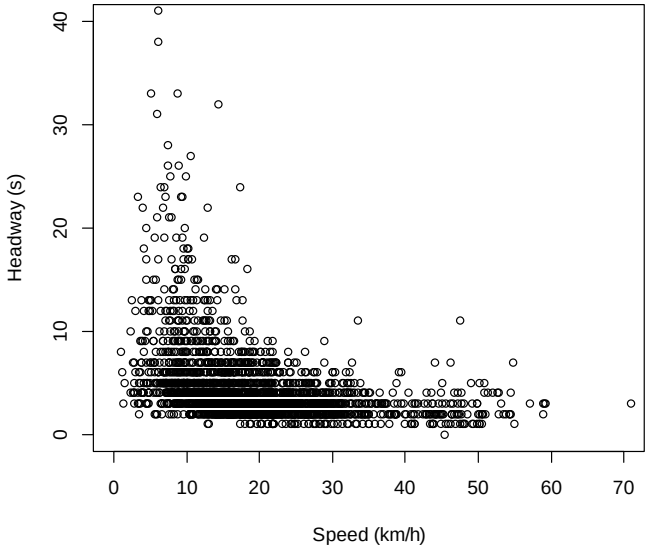
values are above 50 kph 7:00 to 8:00) has a different correlation relationship. As shown in Figure 3.2, the possible explanation is that the relationship between speed and headway can be split into two regimes. The time headway is approximately stable when speed is above 20 kph in the first regime. In the second regime when speed is below 20 kph, the time headway increases significantly as speed decreases. Second, Kendall's tau and Spearman's rho values indicate that there exists a very limited relationship between speed and vehicle length. Third, headway and vehicle length have the strongest dependence during the afternoon peak period (i.e., 16:00 to 19:00). For the copula modeling approach, parameter θ is related to Kendall's tau and it is assumed to be fixed. Thus, this modeling approach cannot capture the varying characteristics of dependence structure between speed and headway. However, under the same traffic condition, the dependence structure among speed, headway and vehicle length is quite stable. In the following section, the traffic data observed under the congested traffic condition (from 16:00 to 19:00) are considered to demonstrate the usefulness of copula methods for constructing bivariate models. This is because the relationship between speed and headway and the influence of vehicle length on headway is more obvious in the car following situation. Figure 6.1 (a), (b), (c) and (d) show the scatter plots of speed, headway and vehicle length for the time period from 16:00 to 19:00. Note that there were 2,360 vehicles observed between 16:00 to 19:00.

Table 6.1 Hourly dependence among speed, headway and vehicle length for the 24-hour period.

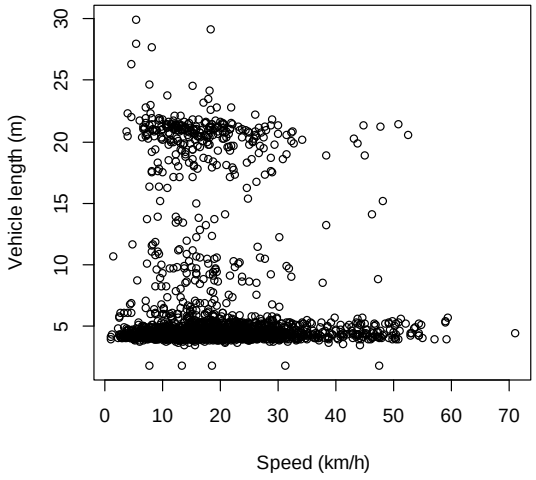
Time period	Count (Vehicles)	Speed & Headway		Speed & Vehicle length		Headway & Vehicle length	
		τ	ρ_S	τ	ρ_S	τ	ρ_S
0 to 1	457	-0.02	-0.03	-0.02	-0.03	-0.02	-0.02
1 to 2	354	-0.01	-0.02	-0.08	-0.11	-0.02	-0.03
2 to 3	301	-0.01	-0.01	-0.12	-0.18	-0.03	-0.04
3 to 4	277	-0.05	-0.08	-0.06	-0.08	-0.03	-0.05
4 to 5	346	-0.02	-0.02	-0.01	-0.01	-0.04	-0.06
5 to 6	709	0.05	0.07	-0.11	-0.16	0.03	0.04
6 to 7	1594	0.15	0.21	-0.02	-0.03	0.08	0.11
7 to 8	2039	0.04	0.05	0.02	0.03	0.10	0.13
8 to 9	1851	0.05	0.07	0.02	0.02	0.09	0.12
9 to 10	1701	0.11	0.15	-0.04	-0.05	0.03	0.04
10 to 11	1653	0.13	0.17	-0.06	-0.10	0.06	0.08
11 to 12	1707	0.10	0.13	-0.03	-0.05	0.09	0.12
12 to 13	1748	0.11	0.15	-0.06	-0.08	0.08	0.11
13 to 14	1739	0.11	0.15	-0.02	-0.03	0.04	0.05
14 to 15	1722	0.12	0.16	-0.01	-0.01	0.11	0.14
15 to 16	1295	-0.35	-0.46	0.00	0.00	0.07	0.09
16 to 17	755	-0.34	-0.45	0.01	0.01	0.14	0.19

Table 6.1 Continued

17 to 18	676	-0.33	-0.45	-0.01	-0.01	0.14	0.19
18 to 19	929	-0.36	-0.49	0.02	0.04	0.13	0.17
19 to 20	1446	-0.11	-0.13	0.01	0.01	0.12	0.16
20 to 21	1241	0.11	0.15	-0.03	-0.04	0.04	0.06
21 to 22	1267	0.08	0.11	-0.01	-0.01	0.05	0.07
22 to 23	1185	0.09	0.12	-0.01	-0.02	0.05	0.07
23 to 24	927	0.03	0.04	-0.03	-0.05	0.01	0.02

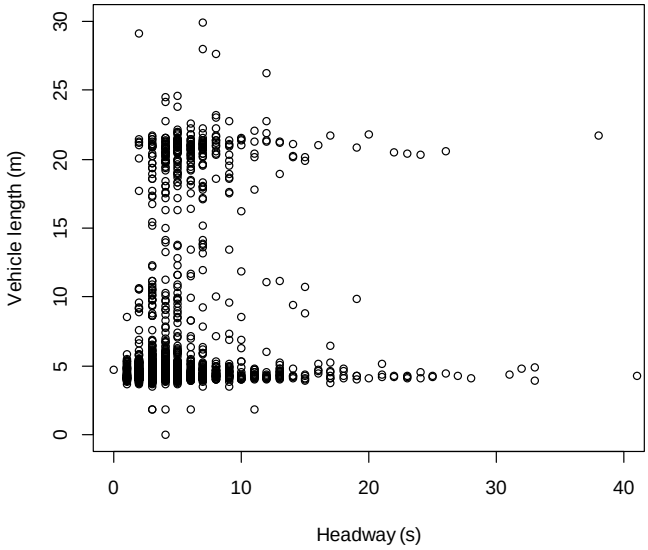


(a)

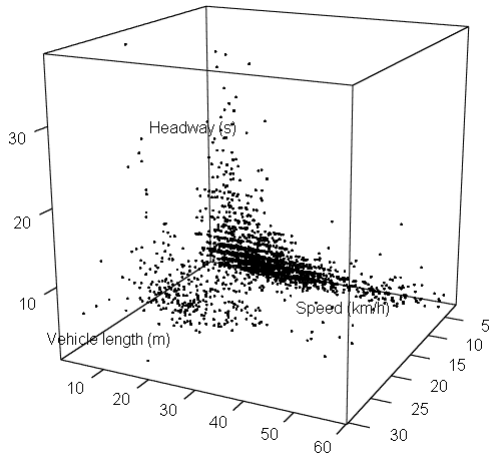


(b)

Figure 6.1 Scatter plot of (a) speed and headway; (b) speed and vehicle length; (c) headway and vehicle length; (d) speed, headway and vehicle length for time period from 16:00 to 19:00.



(c)



(d)

Figure 6.1 Continued

We further examine the dependence among speed, headway and vehicle length using the chi-plot which was proposed by Fisher and Switzer (2001). The chi-plot depends on the data through the values of their ranks and it is defined as follows:

$$H_i = \frac{1}{n-1}\#\left\{ j \neq i : X_j \leq X_i, Y_j \leq Y_i \right\} \tag{6.32}$$

$$F_i = \frac{1}{n-1}\#\left\{ j \neq i : X_j \leq X_i \right\} \tag{6.33}$$

and

$$G_i = \frac{1}{n-1}\#\left\{ j \neq i : Y_j \leq Y_i \right\} \tag{6.34}$$

The above quantities depend exclusively on the ranks of the observations. A chi-plot is a scatter plot of the pairs (λ_i, χ_i), where

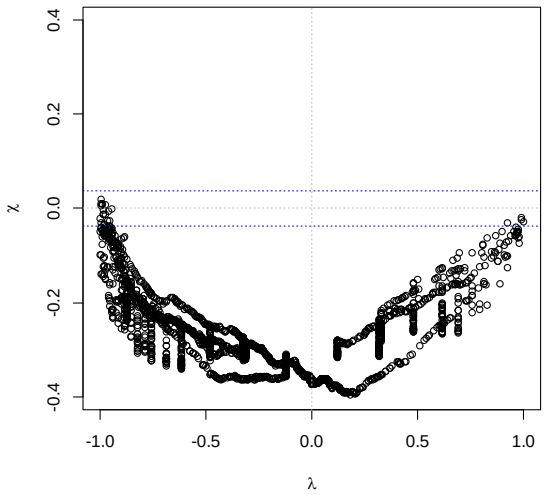
$$\chi_i = \frac{H_i - F_i G_i}{\sqrt{F_i(1-F_i)G_i(1-G_i)}}$$

and

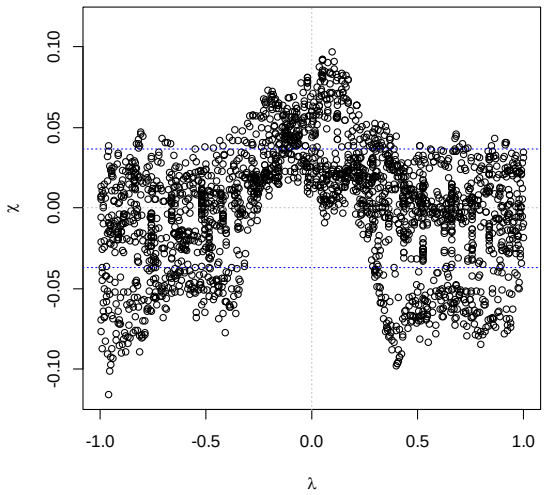
$$\lambda_i = 4\text{sign}\left\{(F_i - 1/2)(G_i - 1/2)\right\}\max\left\{(F_i - 1/2)^2, (G_i - 1/2)^2\right\}.$$

To avoid outliers, Fisher and Switzer (2001) recommend that $|\lambda_i| \leq 4\left(\frac{1}{n-1} - \frac{1}{2}\right)^2$. Figure 6.2 (a), (b) and (c) show the chi-plots for the traffic data observed from 16:00 to 19:00. Dashed blue lines are the 95% confidence band and values of χ_i measure the degree of departures from the hypothesis that speed and headway are independent. As shown in Figure 6.2 (a), almost all points lie below the 95% probability region and this confirms

the presence of negative association between speed and headway. For speed and vehicle length, Figure 6.2 (b) shows that many data points are within the dashed blue lines and the remaining points are either above or below the 95% probability region. Since the area inside the confidence interval means independent, the finding from Figure 6.2 (b) is consistent with the results reported in Table 1 that the evidence in support of the dependence between speed and vehicle length is generally lacking. Figure 6.2 (c) demonstrates that most points are lying above the 95% probability region and while some of the points fall inside the confidence band. This pattern corroborates the presence of positive association between headway and vehicle length.

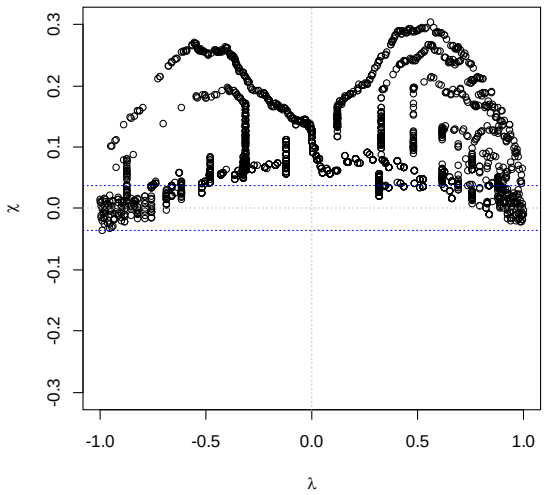


(a)



(b)

Figure 6.2 Chi-plot for (a) speed and headway; (b) speed and vehicle length; (c) headway and vehicle length.



(c)

Figure 6.2 Continued

6.9 Marginal Distribution

In selecting the marginal distributions, we model speed using normal, log-normal, skew-normal and skew-t distributions and headway using gamma, lognormal and log-logistic distributions. Compared with speed and headway, few studies focused on the distribution of vehicle lengths. Previously, Wang and Nihan (2004) and Ye and Zhang (2008) used normal distributions to fit vehicle length data for short and long vehicles. Considering the excess skewness, kurtosis and bimodality present in vehicle length distribution, three mixture models are selected, which are 2-component normal mixture distribution, 2-component skew-normal mixture distribution and 2-component skew-t mixture distribution. The parameters were estimated by the maximum likelihood method. The best fitted distributions for speed, headway and vehicle length were selected using log-

likelihood, the Akaike information criterion (AIC) and root mean square error (RMSE) values. Table 6.2 reports the log-likelihood, AIC and RMSE values of different speed, headway and vehicle length models. Larger log-likelihood and smaller AIC and RMSE values indicate a better overall fit. For the speed data, the skew-t model is better than other models in term of goodness of fit index and normal model provide the least fitting result. In the meantime, the headway data were examined using gamma, lognormal and log-logistic models. The performance of headway models is not consistent. Based on the results, the log-logistic model has the highest log-likelihood and lowest AIC and RMSE values and the gamma model provides the least satisfactory fitting performance. As discussed above, the bimodality of the vehicle length distribution indicates the presence of 2 different clusters. Thus, 2-component mixture distributions were used. The fitting results illustrate that the 2-component skew-t distribution can provide a more accurate description of the bimodal vehicle length distribution than the other two mixture models. Thus, the skew-t, log-logistic and 2-component skew-t distributions are selected as the marginal distributions for describing speed, headway and vehicle length, respectively.

Table 6.2 Log-likelihood, AIC and RMSE values of different fitted probability distributions for each traffic variable

Traffic variable	Fitted marginal distributions	Log-likelihood	AIC	RMSE
Speed	Normal	-8751.15	17506.30	14.75
	Log-normal	-8521.78	17047.56	11.30
	Skew-normal	-8495.50	16997.00	9.69
	Skew-t	-8476.43	16960.86	8.01
Headway	Log-normal	-5250.48	10504.95	23.06
	Gamma	-5488.46	10980.93	48.51
	Log-logistic	-5193.94	10391.89	16.74
Vehicle length	2-component normal	3492.39	6996.78	28.08
	2-component skew-normal	3262.86	6541.72	29.10
	2-component skew-t	3035.33	6090.67	28.30

6.10 Optimal Copula Model Selection

In this section, we modeled the dependence between speed and headway, and headway and vehicle length using different families of copulas. Note that speed and vehicle length are assumed to be independent due to lack of evidence to support the association between each other. The possible explanation is that cars and trucks have the same speed limit on IH-35. The traffic data observed in the congested traffic condition (16:00 to 19:00) were used.

Different copulas introduced in section 6.4 were used and the most appropriate copulas were identified. The calculated Kendall's tau and estimated values of parameter θ of each copula are provided in Table 6.3. Note that Kendall's tau for speed and headway is -0.37. Thus, some copulas can be eliminated immediately, given that the degrees of dependence they span were insufficient to account for the association observed between speed and headway. As a result, only Gaussian and Frank copulas are applicable to the speed and headway data. The best copula model was selected based on log-likelihood, AIC and RMSE values. For speed and headway data, the Gaussian copula can give slightly larger log-likelihood and smaller AIC and RMSE values than the Frank copula. For headway and vehicle length data, all copulas are viable and the goodness-of-fit statistics for each copula model are provided in Table 6.4. Overall, the Gaussian copula was found as the best fitted copula for headway and vehicle length data.

Table 6.3 The estimation of Kendall's tau τ and parameter θ of different copulas

	τ	Gaussian	FGM	Gumbel	Clayton	Ali-Mikhail-Haq	Frank	Joe
Speed and headway	-0.37	-0.55	NA*	NA	NA	NA	-3.80	NA
Headway and length	0.13	0.21	0.59	1.15	0.30	0.51	1.21	1.27

* NA means that the parameter θ for that copula is not applicable. This is because some copulas (Gumbel, Claytion and Joe copulas) can only model positive correlated random variables, i.e., Kendall's tau $\tau > 0$; for the FGM copula, it can model the correlated random variables with $-2/9 \le \tau \le 2/9$; for the Ali-Mikhail-Haq copula, it can model the correlated random variables with $-0.182 < \tau < 0.333$.

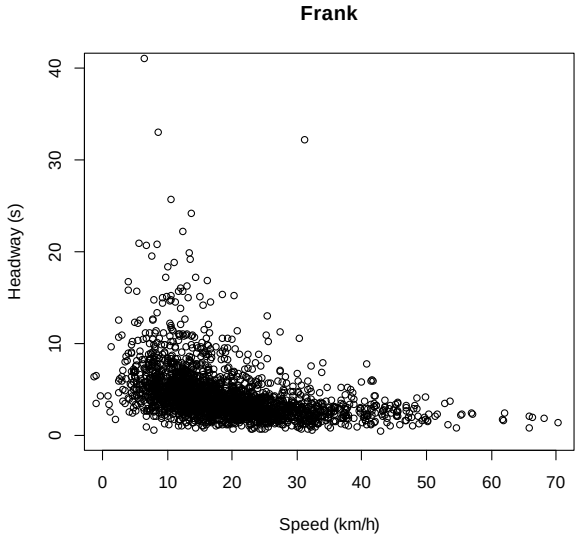
Table 6.4 The log-likelihood, AIC and RMSE values of different copulas

	Goodness-of-fit statistics	Gaussian	FGM	Gumbel	Clayton	Ali-Mikhail-Hq	Frank	Joe
Speed and headway	LL*	-13305.72	NA	NA	NA	NA	-13315.74	NA
	AIC	26625.44	NA	NA	NA	NA	26645.48	NA
	RMSE	1.08	NA	NA	NA	NA	1.12	NA
Headway and length	LL	-8179.75	-8180.72	-8182.55	-8206.01	-8188.79	-8179.55	-8198.61
	AIC	16385.50	16387.44	16391.11	16438.03	16403.59	16385.09	16423.21
	RMSE	4.20	4.27	4.23	4.46	4.32	4.26	4.45

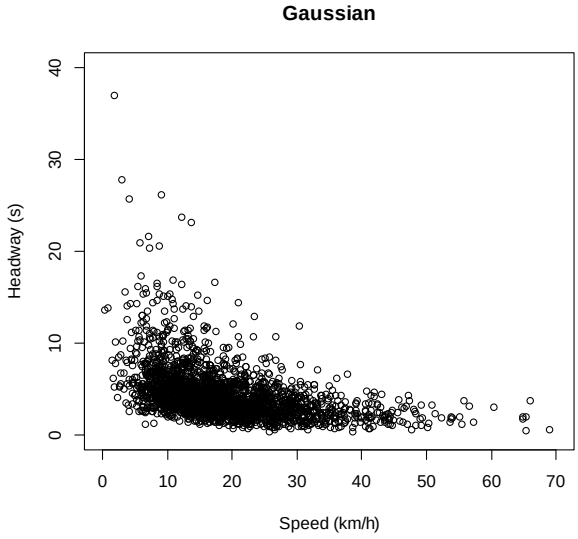
* LL denotes log-likelihood.

One natural way to check the adequacy of copula models is to compare the scatter plot of observations with an artificial dataset of the same size generated from fitted copulas. Using the random variate generation algorithm previously introduced, 2,360 pairs (U_i, V_i) were simulated from the Frank and Gaussian copulas with specified θ values. Then, the 2,360 pairs (U_i, V_i) from each copula model were transformed back into the original units using the marginal distribution identified in the marginal distribution section for speed and headway. Figure 6.3 displays the simulated speed and headway

samples. Assuming $\theta = 0$ for the FGM copula, the independent speed and headway samples were also generated for the purpose of comparison. The actual observations are provided in Figure 6.1 (a). As shown in Figure 6.3 (a) and (b), the simulated samples from the Frank copula and Gaussian copula can accurately reproduce the dependence structure revealed by the speed and headway observations. Moreover, the inappropriateness of the independent model is apparent, as it is hard to observe the inverse relationship between speed and headway from Figure 6.3 (c). The same procedure was repeated for the headway and vehicle length data using various copulas with specified θ values. Figure 6.4 exhibits the simulated headway and vehicle length samples. Due to the very weak dependence between headway and vehicle length, it is hard to tell from Figure 6.4 whether the actual observations can be more accurately reproduced by considering the dependence structure.

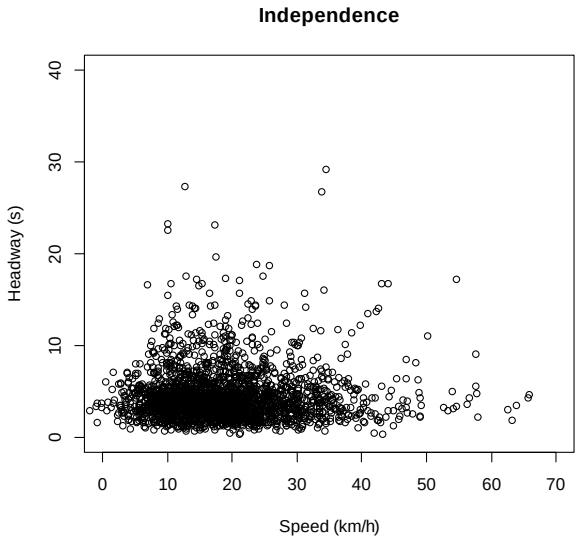


(a)



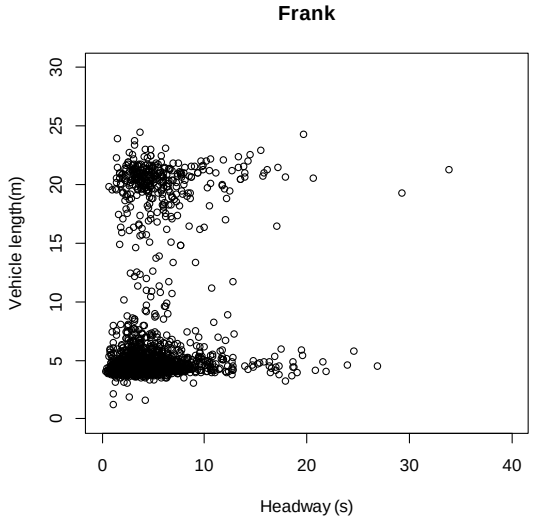
(b)

Figure 6.3 Transformed samples for (a) the Frank copula with parameter $\theta = -3.80$; (b) the Gaussian copula with parameter $\theta = -0.55$; (c) the independent copula.

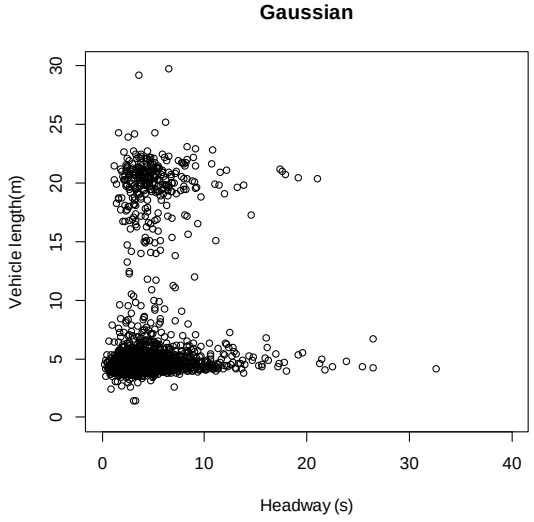


(c)

Figure 6.3 Continued

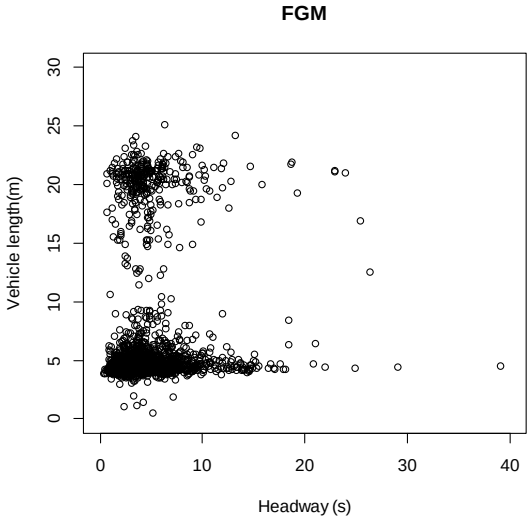


(a)

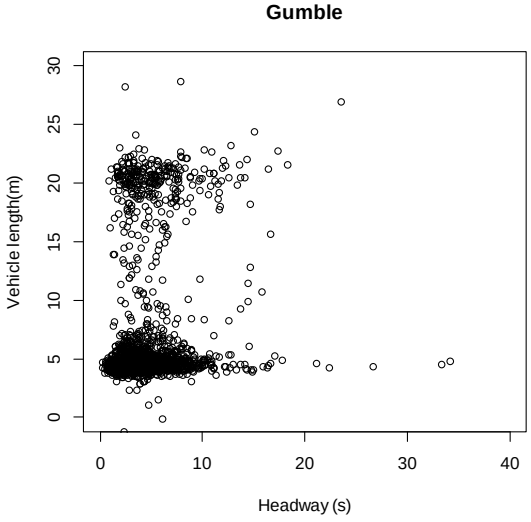


(b)

Figure 6.4 Transformed samples for (a) the Frank copula with parameter $\theta = 1.21$; (b) the Gaussian copula with parameter $\theta = 0.21$; (c) the FGM copula with parameter $\theta = 0.59$; (d) the Gumble copula with parameter $\theta = 1.15$; (e) the Clayton copula with parameter $\theta = 0.3$; (f) the AMH copula with parameter $\theta = 0.51$; (g) the Joe copula with parameter $\theta = 1.27$; (h) the independent copula.

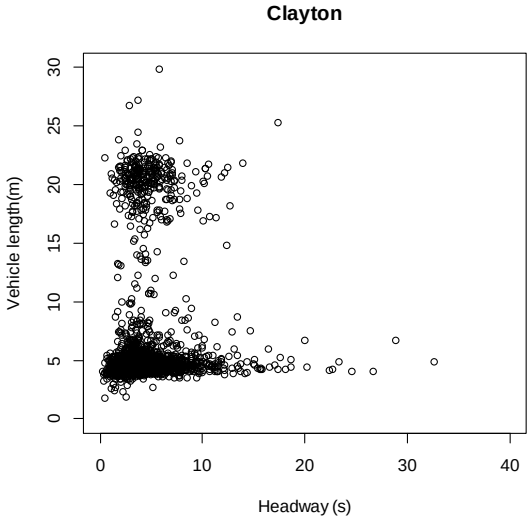


(c)

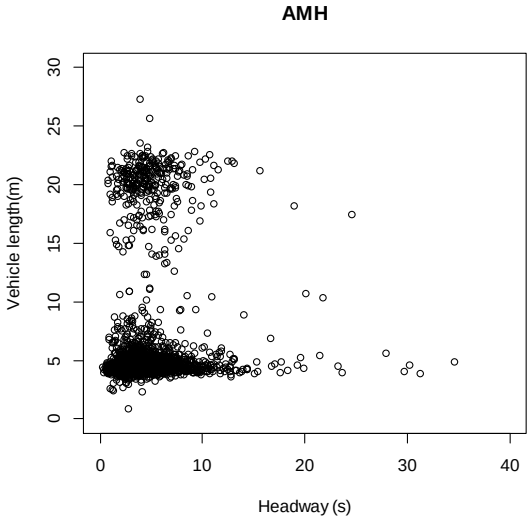


(d)

Figure 6.4 Continued

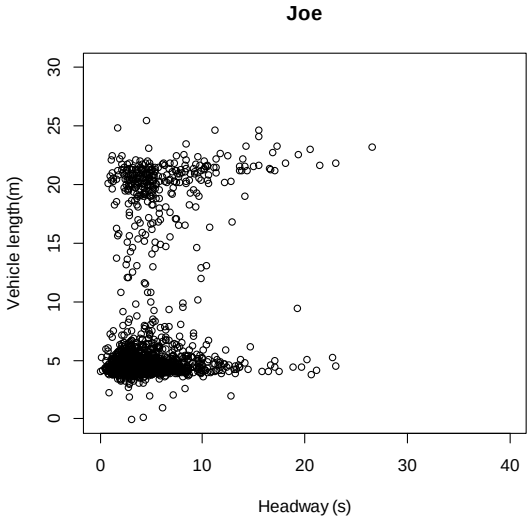


(e)

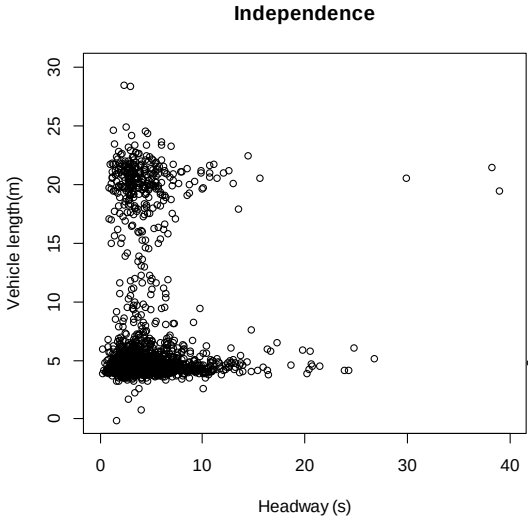


(f)

Figure 6.4 Continued



(g)



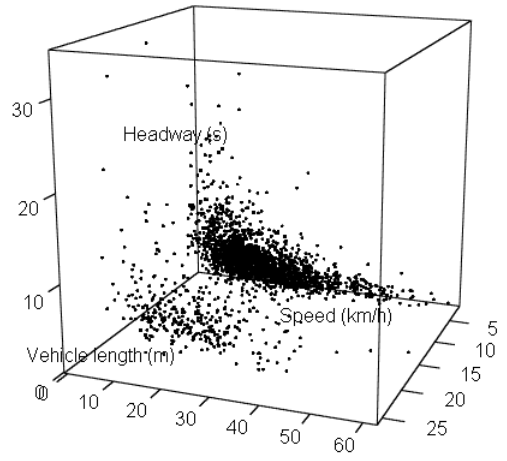
(h)

Figure 6.4 Continued

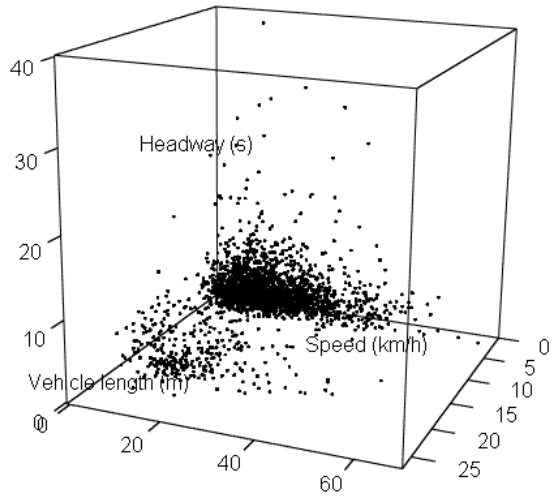
The parameters of trivariate Gaussian copula are estimated and provided in Table 6.5. The log-likelihood, AIC and RMSE values are employed to measure the fitting performance. Using the random variate generation algorithm for the Gaussian copulas, 2,360 vectors $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ were simulated with the specified correlation matrix P. Then, the 2,360 vectors $(\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3)$ were transformed back into the original units using the marginal distribution selected for speed, headway and vehicle length. Figure 6.5 displays the simulated samples. Assuming $\mathrm{P} = I_3$, where I_3 is the 3-dimentional identity matrix, the independent speed, headway and vehicle length samples were also generated for the purpose of comparison. The actual observations are provided in Figure 6.1 (d). Since there is an inverse relationship between speed and headway for both passenger cars and trucks, the simulated samples from the trivariate Gaussian copula can accurately reproduce this dependence structure. However, it is difficult to observe the inverse relationship between speed and headway from Figure 6.5 (b).

Table 6.5 Parameters and fitting evaluation of trivariate Gaussian copula

Parameter			LL	AIC	RMSE
$\rho_{speed\&headway}$	$\rho_{speed\&vehicle\ length}$	$\rho_{headway\&vehicle\ length}$			
-0.55	0.01	0.21	-16306.47	33148.44	0.20



(a)



(b)

Figure 6.5 Transformed samples for (a) the trivariate Gaussian copula; (b) the independent copula.

6.11 Comparison of Copula Models with the Multivariate Skew-t Distribution

In Chapter V, the multivariate skew-t distribution has been applied to the correlated speed and headway data. To compare the performance of the multivariate skew-t distributions with copulas, the traffic data observed in the congested traffic condition (16:00 to 19:00) were analyzed in this section. Considering that vehicle length explicitly consists of two sub-populations (i.e., passenger cars and trucks), the 2-component multivariate skew-t mixture model were used to capture the bimodality of the vehicle length distribution. The probability density function (PDF) of a 2-component mixture of multivariate skew-t distributions is given by

$$f(\mathbf{y} \mid \boldsymbol{\Theta}) = w_1 ST_p(\mathbf{y} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\lambda}_1, \nu) + w_2 ST_p(\mathbf{y} \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\lambda}_2, \nu) \tag{6.35}$$

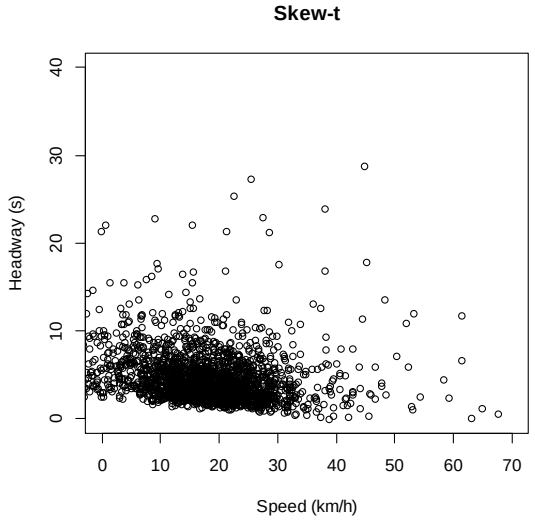
where w_j is the weight of component j, $w_1, w_2 \geq 0$, $w_1 + w_2 = 1$, $\boldsymbol{\Theta} = ((\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \boldsymbol{\lambda}_1, \nu, w_1), (\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2, \boldsymbol{\lambda}_2, \nu, w_2))^T$ is the vector of all parameters.

The multivariate skew-t distributions are applied to the traffic data and the goodness-of-fit statistics are provided in Table 6.6. The copula-based joint distributions and multivariate skew-t distributions were compared using some goodness-of-fit statistics (i.e., the log-likelihood, AIC and RMSE). For the three scenarios, all goodness-of-fit statistics indicate that the copula-based distribution can provide a better fitting performance than the multivariate skew-t distribution and the copula-based joint distribution can describe the distribution of traffic variables more accurately. Three artificial datasets of 2,360 observations were generated from fitted multivariate skew-t distributions and were provided in Figure 6.6. Compared with the actual observations

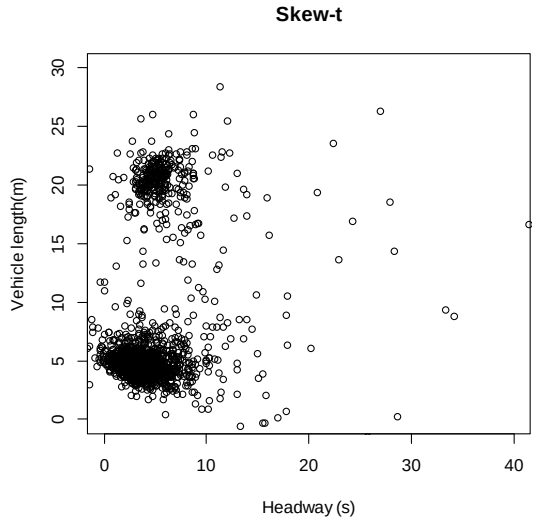
shown in Figure 6.1, the distribution of simulated data points differ significantly from the empirical data. In particular, contrary to the findings in the data analysis section, it is difficult to observe the inverse relationship between speed and headway in Figure 6.6 (a) and (c).

Table 6.6 Fitting evaluation of multivariate skew-t distributions

	Fitted distribution	LL	AIC	RMSE
Speed and headway	bivariate skew-t distribution	-13649.41	27316.82	1.23
Headway and vehicle length	2-component mixture of bivarite skew-t distribution	-8609.02	17240.03	4.73
Speed, headway and vehicle length	2-component mixture of trivariate skew-t distribution	-17435.49	34908.98	0.22

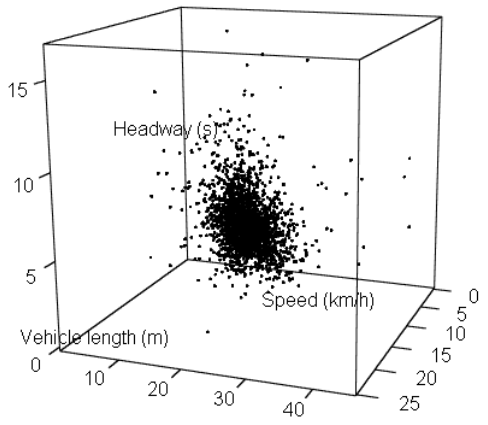


(a)



(b)

Figure 6.6 Simulated samples from multivariate skew-t distributions for (a) speed and headway; (b) headway and vehicle length; (c) speed, headway and vehicle length.



(c)

Figure 6.6 Continued

6.12 Limitation of Copulas

Since the 24-hour traffic data used in this chapter consists of distinct traffic flow conditions, the correlation structure between speed and headway varies based on the traffic condition (for example, as shown in Table 3.1, speed and headway usually have an inverse relationship during the peak period and a positive relationship during the off-peak period.). For copulas, although different marginal distributions can be defined for the one-dimensional speed or headway data, the association parameter θ is assumed to be fixed, which neglects the dynamic nature of the correlation structure between speed and headway over the 24-hour period. The finite mixtures of multivariate distributions can address this issue naturally, since each component has its own covariance matrix and the correlation structure between speed and headway can be different across components.

Thus, when modeling heterogeneous speed and headway data, the finite mixtures of multivariate distributions are preferred over the copula modeling approach.

6.13 Summary

This chapter documented the application of copula models for constructing the distribution of traffic variables (speed, headway and vehicle length) using recorded data collected on IH-35. Before constructing multivariate distributions, we first evaluated the hourly dependence among speed, headway and vehicle length for the 24-hour period. For each hour, Kendall's tau τ , and Spearman's rho ρ_S are used to measure the dependence. Based on the analysis results, the important conclusions can be summarized as follows:

(1) The relationship between speed and headway and the influence of vehicle length on headway is most obvious for the time period from 16:00 to 19:00, which is the busiest time of the day on IH-35.

(2) Vehicle length seems to have a very limited negative effect on vehicle operating speed under both congested and uncongested traffic conditions.

(3) There exists a very weak positive dependence between headway and vehicle length under both congested and uncongested traffic conditions. And vehicle length does influence following headway as trucks and buses usually keep larger following time headways than cars at the same speed level.

After evaluating the dependence among speed, headway and vehicle length, copula models were used to construct bivariate and trivariate traffic distributions and goodness-of-fit statistics showed that the proposed copula models can adequately represent the multivariate distributions of traffic data. Moreover, the simulated samples from some families of copulas can accurately reproduce the actual relationship between traffic variables. Since speed and headway usually have a weak negative correlation under the congested traffic condition, the degrees of dependence most copulas span are insufficient to account for the association. In this chapter, only Gaussian and Frank copulas are applicable to the speed and headway data. Compared with the finite mixtures of multivariate distributions, this chapter shows that copulas can provide better fitting performance and more accurate simulation results. However, since parameter θ is assumed to be fixed, copulas cannot be used to model heterogeneous speed and headway data over an extended period of time with varying traffic conditions. Overall, Chapter VI provides a framework for generating vehicle speeds, vehicle length and vehicle arrival times simultaneously by considering their dependence.

CHAPTER VII

SUMMARY AND CONCLUSIONS

7.1 Summary

Traditionally, traffic variables (speed and headway) are often not studied jointly in microscopic simulation models. One important flaw associated with the traditional approach is that the simulated samples based on the independence assumption usually fail to consider the empirical dependence between traffic variables. To overcome this potential problem associated with the traditional approach, it is necessary to construct bivariate distributions to model vehicle speed and headway simultaneously.

The dissertation first examined the dependence structure between speed and headway using three measures of dependence (i.e., Pearson correlation coefficient, Spearman's rho and Kendall's tau). The dissertation proposed the skew-t mixture models to capture heterogeneity present in speed distribution. To develop a bivariate distribution for capturing the dependence, finite mixtures of multivariate skew-t distributions were applied to the 24-hour speed and headway data. To avoid the restriction of the multivariate skew-t distributions, the dissertation considered copulas as an alternative method for constructing the multivariate distribution of traffic variables.

7.2 Conclusions

Based on the modeling results from this research, we drew some important conclusions, which are listed as follows:

1. The proposed skew-t mixture models can reasonable account for heterogeneity problem in freeway vehicle speed data. Finite mixture of skew-t distributions can significantly improve the goodness of fit of speed data. The methodology developed in this dissertation can be used in analyzing the characteristics of freeway speed data. Considering that many traffic analytical and simulation models use speed as an input for travel time and level of service determination, the developed models can generate more accurate speed value as the input and help improving the reliability of the analysis output.

2. There exists weak dependence between speed and headway and the correlation structure can vary depending on the traffic condition. The dependence between speed and headway is strongest under the most congested traffic condition. Vehicle length seems to have a very limited negative effect on vehicle operating speed under both congested and uncongested traffic conditions. There exists a very weak positive correlation between headway and vehicle length under both congested and uncongested traffic conditions.

3. The bivariate skew-t mixture model can provide a satisfactory fit to the multimodal speed and headway distribution and this modeling approach can accommodate the varying correlation coefficient. For the 24-hour freeway speed and headway data, the three-component bivariate skew-t mixture model was selected as the

optimal model. The proposed methodology can overcome the correlation problem associated with the traditional approach.

4. Copula models can adequately represent the multivariate distributions of microscopic traffic data. Some families of copulas can accurately reproduce the dependence structure revealed by the speed and headway observations. The Gaussian and Frank copulas are applicable to construct the bivariate distribution of speed and headway data with a weak negative dependence. Overall, copula models provide an accurate way for simulating vehicle speeds, vehicle length and vehicle arrival times simultaneously under a given flow condition.

7.3 Future Research

This research proposes two different methodologies to construct bivariate distributions to describe the characteristics of speed and headway, and there are some avenues for future work.

1. A better understanding of speed and headway distributions and its dependence structure can help operational analysis of a freeway facility. In future, since the speed and headway data are site dependent and different sites may have distinct traffic characteristics, multiple locations should be investigated to fully explore the relationship between speed and headway.

2. Traffic headway includes time headway and distance headway, which are closely related to each other and both vary depending on speed and traffic condition. Distance headway is also an important microscopic traffic variable and one influential factor in

the car following model. Some studies have shown that there exists positive dependence between distance headway and speed. Thus, if the distance headway data is available in this study, we can further investigate the dependence structure among distance headway speed, vehicle length. The findings from further analysis may contribute to the existing car following theory.

3. In some popular traffic simulation models (i.e., CORSIM, SimTraffic and VISSIM), vehicles are usually generated on the basis of a certain headway distribution. CORSIM considers three types of vehicle entry headway generation distributions: uniform, normal and Erlang distributions. The negative exponential distribution is used in VISSIM and SimTraffic. The current simulation protocols in these microscopic traffic simulation models fail to consider the dependence between speed and headway. Thus, in the future, the copula-based distributions can be used in these traffic simulation models to generate more accurate speed and headway of entry vehicle.

REFERENCES

ALI, M. M., MIKHAIL, N. & HAQ, M. S. 1978. A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis,* 8**,** 405-412.

AZZALINI, A. 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics,* 12**,** 171-178.

AZZALINI, A. & CAPITANIO, A. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society Series B-Statistical Methodology,* 65**,** 367-389.

AZZALINI, A. & DALLAVALLE, A. 1996. The multivariate skew-normal distribution. *Biometrika,* 83**,** 715-726.

BASSO, R. M., LACHOS, V. H., CABRAL, C. R. B. & GHOSH, P. 2010. Robust mixture modeling based on scale mixtures of skew-normal distributions. *Computational Statistics & Data Analysis,* 54**,** 2926-2941.

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**,** 749-765.

BIERNACKI, C., CELEUX, G. & GOVAERT, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 22**,** 719-725.

BRACKSTONE, M., WATERSON, B. & MCDONALD, M. 2009. Determinants of following headway in congested traffic. *Transportation Research Part F-Traffic Psychology and Behaviour,* 12**,** 131-142.

CABRAL, C. R. B., LACHOS, V. H. & PRATES, M. O. 2012. Multivariate mixture modeling using skew-normal independent distributions. *Computational Statistics & Data Analysis,* 56**,** 126-142.

CHERUBINI, U., LUCIANO, E. & VECCHIATO, W. 2004. *Copula methods in finance*, Wiley, New York.

CLAYTON, D. G. 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika,* 65**,** 141-151.

COWAN, R. J. 1975. Useful headway models. *Transportation Research,* 9**,** 371-375.

DEMARTA, S. & MCNEIL, A. J. 2005. The t copula and related copulas. *International Statistical Review,* 73**,** 111-129.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. 1977. Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society Series B-Methodological,* 39**,** 1-38.

DEY, P. P. & CHANDRA, S. 2009. Desired time gap and time headway in steady-state car-following on two-lane roads. *Journal of Transportation Engineering-Asce,* 135**,** 687-693.

DEY, P. P., CHANDRA, S. & GANGOPADHAYA, S. 2006. Speed distribution curves under mixed traffic conditions. *Journal of Transportation Engineering-Asce,* 132**,** 475-481.

EMBRECHTS, P., LINDSKOG, F. & MCNEIL, A. 2003. Modelling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance,* 8**,** 329-384.

EMBRECHTS, P., MCNEIL, A. & STRAUMANN, D. 2002. Correlation and dependence in risk management: properties and pitfalls. *Risk Management: Value at Risk and Beyond***,** 176-223.

FISHER, N. & SWITZER, P. 2001. Graphical assessment of dependence: Is a picture worth 100 tests? *The American Statistician,* 55**,** 233-239.

FRÜHWIRTH-SCHNATTER, S. 2006. *Finite mixture and Markov switching models*. Springer Series in Statistics. Springer, New York.

FRANK, M. J. 1979. On the simultaneous associativity of F (x, y) and x+y− F (x, y). *Aequationes Mathematicae,* 19**,** 194-226.

GENEST, C. & FAVRE, A. C. 2007. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrologic Engineering,* 12**,** 347-368.

GENEST, C., GHOUDI, K. & RIVEST, L.-P. 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika,* 82**,** 543-552.

GENEST, C. & MACKAY, R. J. 1986. Copules archimédiennes et families de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics,* 14**,** 145-159.

GENEST, C. & RIVEST, L.-P. 1993. Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American statistical Association,* 88**,** 1034-1043.

GERLOUGH, D. L. & HUBER, M. J. 1976. Traffic flow theory. Special Report 165. Transportation Research Board, Washington, DC.

GRIMALDI, S. & SERINALDI, F. 2006. Asymmetric copula in multivariate flood frequency analysis. *Advances in Water Resources,* 29**,** 1155-1167.

GUMBEL, E. J. 1960. Bivariate exponential distributions. *Journal of the American Statistical Association,* 55**,** 698-707.

HAIGHT, F. A. 1965. *Mathematical theories of traffic flow*. Academic Press, New York.

HAIGHT, F. A. & MOSHER, W. W. 1962. A practical method for improving the accuracy of vehicular speed distribution measurements. *Highway Research Board Bulletin*.

HOFERT, M., M CHLER, M. & MCNEIL, A. J. 2012. Likelihood inference for Archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis,* 110**,** 133-150.

HOOGENDOORN, S. P. & BOVY, P. H. 1998. New estimation technique for vehicle-type-specific headway distributions. *Transportation Research Record: Journal of the Transportation Research Board,* 1646**,** 18-28.

JOE, H. 1993. Parametric families of multivariate distributions with given margins. *Journal of Multivariate Analysis,* 46**,** 262-282.

JOE, H. 1997. *Multivariate models and dependence concepts*, Chapman and Hall, London.

JUN, J. 2010. Understanding the variability of speed distributions under mixed traffic conditions caused by holiday traffic. *Transportation Research Part C-Emerging Technologies,* 18**,** 599-610.

KIM, G., SILVAPULLE, M. J. & SILVAPULLE, P. 2007. Comparison of semiparametric and parametric methods for estimating copulas. *Computational Statistics & Data Analysis,* 51**,** 2836-2850.

KO, J. & GUENSLER, R. L. Characterization of congestion based on speed distribution: a statistical approach using Gaussian mixture model. Transportation Research Board Annual Meeting, 2005.

LEE, S. & MCLACHLAN, G. J. 2013. Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing***,** 1-22.

LEONG, H. 1968. The distribution and trend of free speeds on two lane two way rural highways in New South Wales. Australian Road Research Board (ARRB) Conference, 4th, Melbourne.

LIN, T. I. 2010. Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing,* 20**,** 343-356.

LIN, T. I., LEE, J. C. & HSIEH, W. J. 2007. Robust mixture modeling using the skew t distribution. *Statistics and Computing,* 17**,** 81-92.

LUTTINEN, R. 1992. Statistical properties of vehicle time headways. *Transportation research record*.

LUTTINEN, R. T. 1999. Properties of Cowan's M3 headway distribution. *Transportation Research Record: Journal of the Transportation Research Board,* 1678**,** 189-196.

MAY, A. D. 1990. *Traffic Flow Fundamentals*, Prentice Hall, Englewood Cliffs, New Jersey.

MCLEAN, J. Observed speed distributions and rural road traffic operations. Australian Road Research Board Conference Proc, 1979.

NELSEN, R. B. 2006. *An introduction to copulas*, Springer, New York.

PARK, B. J., ZHANG, Y. L. & LORD, D. 2010. Bayesian mixture modeling approach to account for heterogeneity in speed data. *Transportation Research Part B-Methodological,* 44**,** 662-673.

SARVI, M. 2011. Heavy commercial vehicles ‐ following behavior and interactions with different vehicle classes. *Journal of Advanced Transportation*.

SCHUCANY, W. R., PARR, W. C. & BOYER, J. E. 1978. Correlation structure in Farlie-Gumbel-Morgenstern distributions. *Biometrika,* 65**,** 650-653.

SKLAR, M. 1959. *Fonctions de répartition à n dimensions et leurs marges*, Université Paris 8.

TAIEB-MAIMON, M. & SHINAR, D. 2001. Minimum and comfortable driving headways: Reality versus perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society,* 43**,** 159-172.

TONG, Y. L. 1990. Multivariate normal distribution, Springer, New York.

WANG, Y. & NIHAN, N. L. Dynamic estimation of freeway large-truck volumes based on single-loop measurements. Intelligent Transportation Systems, 2004. Taylor & Francis, 133-141.

WASIELEWSKI, P. 1979. Car-following headways on freeways interpreted by the semi-Poisson headway distribution model. *Transportation Science,* 13**,** 36-55.

WINSUM, W. V. & HEINO, A. 1996. Choice of time-headway in car-following and the role of time-to-collision information in braking. *Ergonomics,* 39**,** 579-592.

WYMAN, J. H., BRALEY, G. A. & STEVENS, R. I. 1985. *Field evaluation of FHWA vehicle classification categories*, Maine Department of Transportation, Bureau of Highways, Materials and Research Division.

YE, F. & ZHANG, Y. 2009. Vehicle type-specific headway analysis using freeway traffic data. *Transportation Research Record: Journal of the Transportation Research Board,* 2124**,** 222-230.

YE, Z., ZHANG, Y. & MIDDLETON, D. R. 2006. Unscented Kalman filter method for speed estimation using single loop detector data. *Transportation Research Record: Journal of the Transportation Research Board,* 1968**,** 117-125.

YIN, S. C., LI, Z. H., ZHANG, Y., YAO, D., SU, Y. L. & LI, L. 2009. Headway distribution modeling with regard to traffic status. *2009 IEEE Intelligent Vehicles Symposium, Vols 1 and 2***,** 1057-1062.

ZHANG, G., WANG, Y., WEI, H. & CHEN, Y. 2007. Examining headway distribution models with urban freeway loop event data. *Transportation Research Record: Journal of the Transportation Research Board,* 1999**,** 141-149.

ZHANG, L. & SINGH, V. P. 2006. Bivariate flood frequency analysis using the copula method. *Journal of Hydrologic Engineering,* 11**,** 150-164.

ZHANG, Y., XIE, Y. & YE, Z. Estimation of large truck volume using single loop detector data. Transportation Research Board 87th Annual Meeting, 2008.

ZOU, Y., ZHANG, Y. & LORD, D. 2012. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention,* 50, 1042-1051.