Latent Class Model in Transportation Study

Dengfeng Zhang

Dissertation submitted to the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Feng Guo, Chair

Inyoung Kim

Hesham Rakha

Xinwei Deng

December 3rd, 2014

Blacksburg, Virginia

Keywords: Transportation, Mixture model, Quantile regression

Latent Class Model in Transportation Study

Dengfeng Zhang

(ABSTRACT)

Statistics, as a critical component in transportation research, has been widely used to analyze driver safety, travel time, traffic flow and numerous other problems. Many of these popular topics can be interpreted as to establish the statistical models for the latent structure of data. Over the past several years, the interest in latent class models has continuously increased due to their great potential in solving practical problems. In this dissertation, I developed several latent class models to quantitatively analyze the hidden structure of transportation data and addressed related application issues.

The first model is focused on the uncertainty of travel time, which is critical for assessing the reliability of transportation systems. Travel time is random in nature, and contains substantial variability, especially under congested traffic conditions. A Bayesian mixture model, with the ability to incorporate the influence from covariates such as traffic volume, has been proposed. This model advances the previous multi-state travel time reliability model in which the relationship between response and predictors was lacking.

The Bayesian mixture travel time model, however, lack the power to accurately predict the future travel time. The analysis indicates that the independence assumption, which is difficult to justify in real data, could be a potential issue. Therefore, I proposed a Hidden Markov model to accommodate dependency structure, and the modeling results were significantly improved.

The second and third parts of the dissertation focus on the driver safety identification. Given the demographic information and crash history, the number of crashes, as a type of count data, is commonly modeled by Poisson regression. However, the over-dispersion issue within

the data implies that a single Poisson distribution is insufficient to depict the substantial variability. Poisson mixture model is proposed and applied to identify risky and safe drivers. The lower bound of the estimated misclassification rate is evaluated using the concept of overlap probability. Several theoretical results have been discussed regarding the overlap probability. I also introduced quantile regression based on discrete data to specifically model the high-risk drivers.

In summary, the major objective of my research is to develop latent class methods and explore the hidden structure within the transportation data, and the approaches I employed can also be implemented for similar research questions in other areas.

Acknowledgements

I would like to gratefully and sincerely thank Dr. Feng Guo for his understanding, patience, and most importantly, his friendship during my graduate studies at Virginia Tech. He encourages me to explore the common problems in greater depth and his guidance is also essential for me to establish my long-term career goals.

I would also like to thank the members of my doctoral committee, including Inyoung Kim and Xinwei Deng from the Department of Statistics, and Hesham Rakha from the Department of Civil and Environmental Engineering. They have provided me with a great amount of assistance and without their valuable suggestions my dissertation will never be done.

During the five and a half years life time in Blacksburg, my classmates and friends have helped me in several aspects of the life and study. I have learnt a great deal from them.

Finally, I would like to thank my parents. They have taught me the importance of education since I was a child and their support is always my motivation to venture forward.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistics, as a critical component in transportation research, has been widely used to analyze driver safety, travel time, traffic flow and numerous other problems. Many of these popular topics can be interpreted as to establish the statistical models for the latent structure of data. Over the past several years, the interest in latent class models has continuously increased due to their great potential in solving practical problems.

For example, given a data set containing vehicle drivers' crash history, could we classify them as "safe" or "risky" (Venezian 1981)? Could demographic information, such as gender, age, and annual income be used to evaluate the individual driver risk (Massie et al. 1995)?

To explore the latent structure of the data sets is usually considered as a typical problem of *unsupervised learning*, which adapts behavior without being given responses (*supervised learning*) or without even any hints about the goodness of fitting (*reinforcement learning*)

(Oja 2002). A typical example of unsupervised learning is clustering analysis, whose objective is to group objects in such a way that objects in the same group are more similar (or related) to each other than to those in other groups (Tan 2007).

In this chapter, I would like to provide general backgrounds of several important transportation problems as well as the previous statistical practice by a thorough literature review.

1.1 Travel Time Reliability: Bayesian Mixture Model

Travel time of vehicles contains substantial variability. Federal Highway Administration has formally defined travel time reliability as the "consistency or dependability in travel times, as measured from day-to-day or across different times of day". To understand the nature of travel time reliability will help individuals to make trip decision, and the transportation management will be able to improve the efficiency by finding the bottleneck in the system (Guo et al. 2010).

To quantitatively model travel time uncertainty, some researchers applied single-mode distributions (Emam and Ai-Deek 2006, Tu et al. 2008). A number of candidate distributions have been discussed and compared: lognormal, gamma, Weibull and exponential distributions. However, these approaches overlooked the potential heterogeneity within the data. Thus the single-mode distributions usually yield poor model fitting under complex travel conditions, especially during peak hours of a day (Guo et al. 2012).

Due to the lack of flexibility of single-mode distributions, the multi-state models (Fowlkes

1979), or mixture distributions, have drawn researchers' attentions. Not only considerably improving goodness of fit, the multi-state model can also provide the probability that an observation falls in free-flow or congested states. Mixture normal (Guo et al. 2012) and mixture lognormal (Guo et al. 2010) have been applied to the field data collected along a section of the I-35 freeway in San Antonio, Texas (hereafter *I-35 data*). The study covers a sixteen kilometer section with an average daily traffic volume around 150,000 vehicles. The travel time was collected when vehicles tagged by a radio frequency device passed the identification stations on New Braunfels Ave. (Station no. 42) and OConnor Rd. (Station no. 49). The plot below illustrates the data collection procedure.

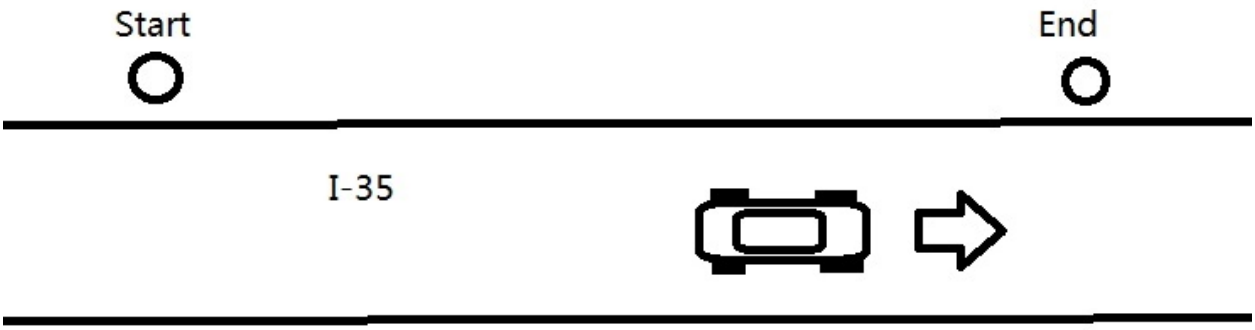

Figure 1.1: Illustration of Data Collection

One of the most attractive features of the multi-state model is its capability to associate

the travel time with underlying traffic conditions. Park et al. (2010) showed that travel time states are related to the fundamental diagram, i.e., traffic flow, speed, and density. Besides the free-flow and congested states, the model can also accommodate delay caused by traffic incidents (Park et al. (2011)). However, one of the most important factors affecting travel time, the *traffic volume* has yet to attract enough attentions from the research community.

The traffic volume, defined as the number of vehicles travel through a specific segment of the road within a specific time period, plays an essential role in my research. I extended the multi-state travel time model by incoporating the effects from traffic volume.

We set the time period to summarize the traffic volume as one hour, and collected the traffic volume from [0:00, 0:59] to [23:00, 23:59] for more than twenty days. According to multi-state travel time model, there are inherently two latent states of traffic condition: "congested" and "free-flow". We applied a regression model on the proportion of each state using the traffic volume as predictor. Based on the work of Guo et al. (2012), the travel time in congested state contained substantial variability, so regression model has also been implemented to the mean of the travel time under such condition.

In sum, two levels of uncertainty are quantitatively assessed in the proposed model. The first level of uncertainty is the probability of a given traffic condition, for example, congesteed or free-flow; the second level of uncertainty is the variation of travel time within each traffic condition.

We proposed a Bayesian model based on Markov Chain Monte Carlo algorithms, and hence we were able to obtain the parameter estimates as well as the uncertainty of esch estimate (Lenk and DeSarbo 2000). The probit function is more convenient in Bayesian context compared to logit function because the corresponding Gibbs sampler is easier to implement (John and Michael 1997).

1.2 Travel Time Reliability: Hidden Markov Model

This chapter is an extension to the previous one. The existing multi-state model is based on the assumption that all the observations are conditionally independent, which might be violated in actual data. It is possible to include auto-correlated structure in the model (Cochrane and Orcutt 1949). An alternative way is to apply hidden Markov model (Baum and Petrie (1966).

Hidden Markov model can be seen as a mixture model which relaxes the independence assumption (Yuting et al. 2007). It is able to incorporate the dependency structure of observations, and also include traditional mixture model as a special case (Scott 2002).

Hidden Markov models have become important in a wide variety of applications including: speech recognition (Rabiner 1989), biometrics (Albert 1991), econometrics (Hamilton 1989), computational biology (Krogh et al. 1994) and etc. We showed that the hidden Markov model can also be used to in travel time data and the result is superior to traditional mixture model.

1.3 Individual Driver Risk: Poisson Mixture and Overlap Probability

Traditional traffic safety research focuses on the influence of environmental factors, such as intersection design features, pavement conditions, weather and traffic flow conditions (Maze et al. 2006).

The traffic safety research community began to study individual driver risk by implementing accident-count based models during the late 1980s and expanded these applications during the last two decades (Lord and Mannering 2010). The most popular models include Poisson and negative binomial (NB) models (Jovanis and Chang 1986, E. Hauer and Lovell 1988, Poch and Mannering 1996).

Poisson model has a limitation that the mean and the variance are equal. The observed accident-count data usually violate this assumption (over-dispersion). Hauer (2001) reported that the over-dispersion observed in crash data could be derive from the measurement errors or the unrepresented traits. Lord and Bonneson (2005) claimed that the over-dispersion arised from the nature of the crash process, which is the result of independent Bernoulli trials when the probability is unequal.

To handle the over-dispersion issue, NB model was proposed (Bliss and Fisher 1953). Many studies have focused on the structure of the dispersion parameter, and some researchers suggested to use a flexible dispersion parameter, which is a function of the covariates (Heydecker

and Wu 2001). However, finding appropriate covariates that influence the over-dispersion can be problematic (Park and Lord 2009) and for heavily over-dispersed data NB model might not perform well (Guo and Trivedi 2002).

An alternative method for over-dispersion is quasi-Poisson model (Wedderburn 1974). In quasi-Poisson model, a dispersion parameter, or scale parameter ϕ is introduced into the relationship between the variance and the mean for either over-dispersion or under-dispersion: $Var(x_i) = \phi E(x_i)$. Ver Hoef and Boveng (2007) compared NB and quasi-Poisson and showed that the variance of a quasi-Poisson model is a linear function of the mean, while the variance of a NB model is a quadratic function of the mean.

One disadvantage of NB and quasi-Poisson model is that a single structure is applied to the entire data set. It is worth noting that the NB distribution can be seen as an infinite mixture of Poisson distributions with Gamma-distributed parameters. In order to identify the latent classes among the data set, some clustering technique, such as K-means has to be applied additionally. Figueiredo and Jain (2002) noted that finite mixture model might be a more coherent solution.

Mixture model has been introduced in the statistics research field for several decades (Teicher 1960). It was also applied in the analysis of the count data (Duijn and Bockenholt, 1995). A finite mixture distribution is a distribution whose sample can be interpreted as being derived from a finite set of classes. The finite mixture model can be described as three parts: the number of components, the proportion of each component and the parameter(s) in each component. For example, suppose all the drivers can be split into risky and safe

groups, and the numbers of crashes of these two groups follow two Possion distributions with different parameters, then the number of crashes for the entire population can be modeled as a mixture of two Poisson components.

It is well known that the traffic accident is rare event for individual driver (Ng et al. 2002, Formisano et al. 2005). For example, Findley et al. (2000) found that the automobile crash rate for all drivers in the state of Colorado per year is no more than 0.1. Therefore, the crash history data in a single year could contain a large number of zeros, which leads to the zero-inflated model (Lambert 1992).

In the zero-inflated model, one of the component is assumed to be point mass at constant zero, and hence excessive zeros in the data set could be handled (Lord et al., 2005). In general, the zeros generated by this model can be decomposed as structural zeros and sampling zeros (Jansakul and Hinde 2002, Mohri and Roark 2005). Chang and Kim (2012) proposed a mixture model of zero-inflated and ordinary Poisson. Lim et al. (2013) generalized this idea and proposed a finite mixture of zero-inflated Poisson. Ghosh et al. (2012) also tried to extend the zero-inflated model in a more flexible framework, which turns out to be a special case of the model established by Baetschmann and Winkelmann (2013).

However, the underlying data generating process of zero-inflated model is controversial in modeling vehicle crash data and the assumption might be too restricted, because even the safest driver should not be guaranteed to be always safe (Lord et al., 2007). Warton (2005) argued that for many real data sets with excessive zeros, taken for granted that a zero-inflated model was required seemed to be a doubtful statement, and negative binomial probability

model was sufficient to handle most occurrences of zero-inflation in environmental and ecological data. Cameron and Trivedi (2013) also noted that zero-inflated model could be seen as a special case of finite mixture Poisson (Cameron and Trivedi 2013). Thus, the common finite mixture Poisson model is still be preferred (Park and Lord, 2009).

To estimate the parameters in mixture model, several approaches have been discussed. One of the oldest methods is based on the moments (Harter 1975, Church and Gale 1995). The method of moments is easy to implement but has no gurantee for optimality (neither locally nor globally).

A more popular approach is expectationmaximization (EM) algorithm. The EM algorithm (Dempster et al. 1977) is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models. EM algorithm can be easily implemented in estimating the mixture model (Hasselblad 1969). Although efficient in computing, EM algorithm is not guaranteed to attain the global maxima (Wu 1983, Archambeau et al. 2003), and it might be sensitive to initial values (Karlis and Xekalaki 2003). Thus, it is recommended that several EM iterations should be tried with different starting points which represents the parameter space.

Some other adjustments have also been proposed for EM algorithm. For example, Grn and Leisch (2004) recommended to use bootstrap on the original data in addition to using different initializations. It has been shown that starting from the solution of a K-means type algorithm can also be an effective strategy (Biernacki et al. 2003, Rau et al. 2011). Wang et al. (1996) also discussed the mixture model with covariate dependent parameters based

on EM algorithm. However, according to Frhwirth-Schnatter (2006), the EM algorithm has several drawbacks, and one of them is that it can fail to converge when the sample size is small or the components are not well separated.

Another important type of estimation methods is based on Bayesian statistics. Bayesian approaches to mixture modelling allows for the hierarchical descriptions of both local and global features of the model (Vounatsou et al. 1998). This framework also allows a complicated structure of the mixture model to be decomposed into a set of simpler structures through the use of hidden or latent variables (Marin et al. 2005).

For Poisson mixture, several researchers have studied Bayesian framework. Viallefont et al. (2002) discussed about different prior specifications as well as the homogeneous or heterogeneous covariate effects. Some relevant topics also attract the researchers' attentions, such as the number of components in a mixture model. Classical statisticians usually handles this issue with hypothesis testing (Lo et al. 2001, Chen and Cheng 1997), while Bayesian statistician applies Dirichlet process prior for unknown number of components (MacEachern and Mller 1998) and reversible jump algorithm for dynamically increasing or decreasing the number of components (Green 1995, Zhang et al. 2004, Dellaportas et al. 1997). Stephens (2000) provided an alternative method rather than reversible jump, Markov birth-death process with proper stationary distribution.

For the crash data provided by the CEI Group Inc. the average number of crashes for each driver is around 0.3 per year, which implies that the mean parameters in the Poisson distribution might be small. When the two mean parameters are too close in a mixture

Poisson model, the estimation becomes highly unreliable. Moreover, the simulation study also indicates that the misclassification rates might be considerable. We proposed an intuitive and effective measurement, derived from the overlap of the two components to assess the lower bound of misclassification rates and discussed how it varies when the parameters changes.

1.4 High Risk Drivers: Quantile Regression for Counts

In traffic safety research, one of the most important problems is to identify high-risk drivers. The insurance industry has a long history of research assessing individual driving risk (Sloan et al. 1995, Willett 1901). The interest of the insurance industry focuses on the cost of accidents and the calculation of premiums based on the varied probability of drivers being involved in accidents.

The focus of risk assessment efforts was on classification of good and bad drivers to facilitate the underwriting and pricing (Venezian 1981). From the perspectives of insurance companies, one crucial requirement is that insurance rates should not be unfairly discriminatory. This leads to the issues of risk classification standards. Walters (1981) summarized the standards in three broad categories: homogeneous, well defined, and practical.

The insurance/actuarial field has recognized the randomness associated with accidents long before traffic engineers (Weber 1970). Driver classification has been used to identify high- and low-risk drivers (Harrington and Doerpinghaus 1993). In addition to the predictive

power of the driver risk prediction models, the cost associated with data collection is an issue. There is a tradeoff between the accuracy level and cost of collecting data. The key is to balance the two parts such that with reasonable data collection cost, a relatively high accuracy level is achieved. Consequently, the use of low-cost variables such as age, gender and territory were considered effective in many risk assessment circumstances (Turner and McClure 2003).

However, the relationship between factors and accidents is not simple. There are contradictory conclusions drawn from different researchers. Monrrez-Espino et al. (2006) claimed that the number of crashes of men was almost twice than that of women in all ages. A longitudinal study conducted in New Zealand identified the following risk factors for driving: male, part-time work, rural residence, marijuana use, and early motorcycle riding (Reeder et al. 1998). However, Laapotti et al. (2003) showed that female drivers were less involved in accidents and committed less traffic offenses. Massie et al. (1995) showed that men had a greater fatality crash risk than women, while women had greater rates of involvement in injury crashes and all police-reported crashes. Li et al. (1998) also concluded that men are more likely to be involved in fatal crash, but when exposure is considered female drivers are not safer than male counterparts. When age is considered to be a significant factor related to traffic safety, some research indicated that teen and older drivers are generally at greater risk than other age groups (Chen et al. 2000, Keating 2007). Some other studies indicated that old drivers do not have a high crash risk, because drivers travelling less (typical for most old drivers) will tend to have higher crash rates than those driving more (Langford et al.

2006, Hakamies-Blomqvist et al. 2002).

The contradictions might derive from sample randomness (Bearden et al. 1982), interactions among causal factors (Jaccard and Turrisi 2003) and the inherent heterogeneity of the data. For the first two issues, there have been numerous literatures. However, the inherent heterogeneity was rarely considered.

Given a data set of crash history and demographic information, it is easy to establish a statistical model for crash counts, such as Poisson and negative binomial regression. However, these models only consider the variation of conditional means across different values of the explanatory variables. If the homogeneity assumption holds, these models serve as decent predictive models for high-risk drivers, represented on the upper tail. However, in the situation that heterogeneity is present the conditional means might lead to biased result (Qin et al. 2010). For example, some aforementioned research claimed that old drivers in general commit more crashes. On the other hand, it is also shown that old drivers are less aggressive (Lajunen and Parker 2001), which could prevent them from being extremely high-risk drivers. In order to gain insights regarding how covariates affect high-risk drivers, it is critical to look not only at the mean but also at various quantiles.

Quantile regression is a statistical technique to estimate and conduct inference for the conditional quantiles, such as median, lower and upper quartiles (Koenker and Bassett 1978). It is particularly useful when the data contain subgroups. It has been widely used as a standard tool in economics to study determinents of wages, discrimination effects, market sales, hedge fund and trends in income inequality (Koenker 2006, Taylor 2007, Buchinsky

1994, Baur and Lucey 2010, Schaeck 2008). The quantile regression has also been applied in ecology (Cade and Noon 2003), public health (Winkelmann 2006) education (Eide and Showalter 1998), and etc. The credit card issuer applies the quantile regression to the credit scoring database to assess the risk factors for consumer loan default (Whittaker et al. 2005), which is quite similar to the driver risk problems raised by insurance companies.

There is no explicit solution to estimate parameters in quantile regression since the loss function in quantile regression is not differentiable at the origin. Therefore, several computational methods have been proposed. Traditional linear programming algorithms, such as simplex method (Barrodale and Roberts 1973) and interior point algorithm (Koenker and Park 1996) were applied. The confidence interval can be established by bootstrap method (Jinyong 1995, Koenker 1994). Bayesian method can be applied through asymmetric Laplace distribution (Yu and Moyeed 2001), and alternative framework were discussed by Kozumi and Kobayashi (2011) and Lancaster and Jae Jun (2010).

Currently almost all the techniques in standard regression have been ported to quantile regression. The semiparametric/nonparametric approaches always receive a lot of attentions. Yu and Jones (1998) applied local-linear method, as well as piecewise polynomials and reversible jump MCMC (Yu 2002). Dabrowska (1992) considered the censored data. Kottas and Krnjaji (2009) and Taddy and Kottas (2010) considered Bayesian approach for semiparametric/nonparametric quantile regression. Yue and Rue (2011), Horowitz and Lee (2005) and Yu (2002) discussed about the additive models. Farcomeni (2012) and Geraci and Bottai (2007) proposed the longitudinal data models. Variable selection, as well as in

standard regression, attracts great attention in the study of quantile regression (Wu and Liu 2009). Bayesian least absolute shrinkage and selection operator (lasso) technique has been discussed by Li et al. (2010) and Alhamzawi et al. (2012) considered Bayesian adaptive lasso. Stochastic search variable selection has been discussed by Yu et al. (2013).

There are also some topics specifically derived from quantile regression. For example, by intuition the 95th percentile should be no smaller than 90th percentile, but it is not guaranteed by the common model estimation method. Bondell et al. (2010) and Yu and Jones (1998) discussed about the non-crossing quantile regression. Another interesting topic is based on the fact that traditional quantile regression aims for the conditional quantiles given the covariates, but Firpo et al. (2009) considered a conceptually different model as "unconditional quantiles". When the p is close to 0 or 1 for p-quantile, extremal quantile regression has been considered by Chernozhukov (2005).

Transportation research still has not fully embraced quantile regression. To hsnfle with the common heterogeneity issue, quantile regression could be a useful tool in this area. In the study of driver safety, one of the major data sources is the crash history, whose observations are clearly discrete and nonnegative. For such type of data, Machado and Silva (2005) proposed "jittering" technique. This technique has been applied by Li and Carriquiry (2005) in the road safety study. Qin and Reyes (2011) and Qin et al. (2010) also applied this method to the crash count data.

We applied quantile regression to assess the high risk drivers according to the crash data. We advanced the research of quantile regression in count data by applying smoothed check

function, derived from the idea of M-estimator. This method can be seen as an alternative to the "jittering" method.

Chapter 2

Travel Time Reliability: Bayesian Mixture Model

2.1 Introduction

Travel time of vehicles contains substantial variability. Federal Highway Administration has formally defined travel time reliability as the "consistency or dependability in travel times, as measured from day-to-day or across different times of day". To understand the nature of travel time reliability will help individuals to make trip decision, and the transportation management will be able to improve the efficiency by finding the bottleneck in the system (Guo et al. 2010).

The multi-state modelhas been developed for modeling travel time reliability and one of

the most attractive features is its capability to associate the travel time with underlying traffic conditions. In the Gaussian mixture model, the travel time y is assumed to follow a two-component mixture distribution with density function:

$$f(y|\lambda, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \lambda f_N(y|\mu_1, \sigma_1^2) + (1-\lambda) f_N(y|\mu_2, \sigma_2^2)$$

where f_N represents the density function of a normal distribution with mean μ_i and variance σ_i^2. Without loss of generality, we will assume that $\mu_1 < \mu_2$. Therefore, λ and $1-\lambda$ are the probability of free-flow state and congested state. μ_1 and μ_2 indicate the mean travel time under free-flow state and congested state. σ_1^2 and σ_2^2 represent the variance of travel time under free-flow state and congested state.

The traffic volume is defined as the number of vehicles travel through a specific segment of the road within some time period. We set the time period as one hour, and calculate the traffic volume from [0:00, 0:59] to [23:00, 23:59] for each day.

The probability of free-flow state, denoted by λ, has support in $(0, 1)$. To link λ with traffic volume x, a common approach is to apply logit function of λ:

$$log(\frac{\lambda}{1-\lambda}) = \beta_0 + \beta_1 * x, \text{ or more general,}$$
$$log(\frac{\lambda}{1-\lambda}) = X\beta, \text{where matrix X contains 1's as the first column.}$$

An alternative is probit function, which is the inverse of standard normal cummulative distribution function:

$$\Phi^{-1}(1-\lambda) = X\beta$$

For Bayesian models, the probit function is preferred due to its ease in Markov Chain Monte Carlo (MCMC). In probit modell, a latent variable $w_i \in \mathbf{R}$ is introduced for each observation to indicate which group this observation belongs to:

$$y_i \in \begin{cases} Group_1 & \text{if } w_i < 0 \\ Group_2 & \text{otherwise} \end{cases}$$

Assume the latent variable $w_i \sim N(X_i\beta, 1)$, where X_i is the i^{th} row in matrix X. It can be shown that:

$$\lambda = 1 - \Phi(X\beta) = P(w < 0 | \mu = X\beta, \sigma^2 = 1)$$

This setting establishs the relationship between proportion of two groups and the covariate. The likelihood function is correspondingly $f_N(y|\mu_1, \sigma_1^2)^{I(w<0)} f_N(y|\mu_2, \sigma_2^2)^{I(w\geq 0)}$.

As shown by Guo et al. (2012), the variability of μ_2 can be substantial. From an engineering perspective, there exists certain relationships between μ_2 and the traffic volume x_i. We proposed two possible models to relate μ_2 with traffic volume:

$$(1)\mu_{2i} = \theta_0 + \theta_1 * x_i = X_i\theta$$

$$(2)\mu_{2i} = \theta_s * \mu_1 + \theta * x_i$$

The first model assumes that μ_1 and μ_2 are estimated individually. The second model assumes that the intercept is proportional to μ_1 with a perdetermined scale parameter θ_s. The motivation of the second model is to ensure the two components are well separated.

Following the convention of Bayesian approach, we use the precision parameter ψ_j to denote the inverse of the variance of the two components (i.e. $1/\sigma_j^2$, j=1,2).

In sum, two levels of uncertainty are quantitatively assessed in the proposed model. The first level of uncertainty is the probability of a given traffic condition, for example, congested or free-flow; the second level of uncertainty is the variation of travel time for each traffic condition.

2.2 Markov Chain Monte Carlo Algorithm

Markov Chain Monte Carlo (MCMC) algorithm was used to draw samples from the joint posterior distribution of the parameters:

$$f(\mu_1, \psi, \beta, \theta, w|X, y) \propto f(y|\mu_1, \psi, \beta, \theta, w, X) f(\mu_1, \psi, \beta, \theta, w|X)$$

$$\propto f(y|\mu_1, \psi, w, \theta) f(w|X, \beta) f(\mu_1, \psi, \beta, \theta|X)$$

$$\propto f(y|\mu_1, \psi, w, \theta) f(w|X, \beta) \pi(\mu_1) \pi(\psi_1) \pi(\psi_2) \pi(\beta) \pi(\theta)$$

$f(y|\mu_1, \psi, w, \theta)$ is the density functin of multi-state normal: $f_N(y|\mu_1, 1/\psi_1)^{I(w<0)} f_N(y|X\theta, 1/\psi_2)^{I(w\geq 0)}$. $f(w|X, \beta)$ is the multivariate normal with mean $X\beta$ and covariance matrix $\mathbf{I}$.

According to Yang and Berger (1996), following non-informative priors can be used in Gaussian distributions:

$$\pi(\mu_1) \propto 1, \pi(\beta_0) \propto 1, \pi(\beta_1) \propto 1, \pi(\theta_0) \propto 1, \pi(\theta_1) \propto 1, \pi(\psi_1) \propto 1/\psi_1, \pi(\psi_2) \propto 1/\psi_2$$

It is desirable that a Bayesian model is not sensitive to the choice of prior distributions. We have tried a series of other priors, such as the normal distribution with different variance (Table 2.1). Due to the large sample size, the results are not significantly influenced and they

are quite similar to that from non-informative priors. Therefore, we stick to non-informative priors in the model.

Table 2.1: Variance of Priors

σ_β^2	∞	10	100	1000
σ_θ^2	∞	10	100	1000

2.2.1 Model 1

We obtain the full conditional distribution for each parameter:

1 The full conditional for w:

$$f(w|...) \propto \prod_{i=1}^{n}(f_N(y_i|\mu_1, 1/\psi_1)I(w_i < 0) + f_N(y_i|X_i\theta, 1/\psi_2)I(w_i \geq 0))f_N(w_i|X_i\beta, 1)$$

This is the multi-state truncated normal. Define $a = f_N(y_i|\mu_1, 1/\psi_1), b = f_N(y_i|X_i\theta, 1/\psi_2)$, then with probability $\frac{a}{a+b}$, w_i is sampled from $f_N(w_i|X_i\beta, 1)$ truncated at $w_i < 0$; with probability $\frac{b}{a+b}$, w_i is sampled from $f_N(w_i|X_i\beta, 1)$ truncated at $w_i \geq 0$

2 The full conditional for μ_1:

$$f(\mu_1|...) \propto \prod_{i=1}^{n}(f_N(y_i|\mu_1, 1/\psi_1)I(w_i < 0) + f_N(y_i|X_i\theta, 1/\psi_2)I(w_i \geq 0))$$

$$\propto \prod_{i:w_i<0} f_N(y_i|\mu_1, 1/\psi_1)$$

$$\sim N(\sum_{i:w_i<0} \frac{y_i}{n_1}, \frac{1}{n_1\psi_1})$$

This is a univariate normal distribution. n_1 is the number of $w_i's$ that are smaller than 0. Corresponding to the model assumption $\mu_1 < \mu_2$, we will right truncate this distribution at $min(X_i\theta)$

3 The full conditional for ψ_1:

$$f(\psi_1|...) \propto \psi_1^{-1} \prod_{i=1}^{n} f_N(y_i|\mu_1, 1/\psi_1)^{I(w_i<0)} f_N(y_i|X_i\theta, 1/\psi_2)^{I(w_i\geq 0)}$$
$$\sim \psi_1^{\frac{n_1}{2}-1} exp(-\frac{1}{2}\psi_1 \sum_{i:w_i<0} (y_i - \mu_1)^2)$$

This is the Gamma distribution with shape parameter $\frac{n_1}{2}$ and rate parameter $\frac{1}{2}\sum_{i:w_i<0}(y_i - \mu_1)^2$.

4 The full conditional for ψ_2:

$$f(\psi_2|...) \propto \psi_2^{-1} \prod_{i=1}^{n} f_N(y_i|\mu_1, 1/\psi_1)^{I(w_i<0)} f_N(y_i|X_i\theta, 1/\psi_2)^{I(w_i\geq 0)}$$
$$\sim \psi_2^{\frac{n_2}{2}-1} exp(-\frac{1}{2}\psi_2 \sum_{i:w_i\geq 0} (y_i - X_i\theta)^2)$$

n_2 is the number of $w_i's$ that are greater than or equal to 0. This is the Gamma distribution with shape parameter $\frac{n_2}{2}$ and rate parameter $\frac{1}{2}\sum_{i:w_i\geq 0}(y_i - X_i\theta)^2$.

5 The full conditional for β:

$$f(\beta|...) \propto \prod_{i=1}^{n} f(w_i|X_i, \beta)$$
$$\propto exp(-\frac{\sum_{i=1}^{n}(w_i - X_i\beta)^2}{2})$$

This is the bivariate normal distribution with mean $(X^TX)^{-1}X^Tw$ and covariance matrix $(X^TX)^{-1}$.

6 The full conditional for θ:

$$f(\theta|...) \propto \prod_{i=1}^{n} f_N(y_i|\mu_1, 1/\psi_1)^{I(w_i<0)} f_N(y_i|X_i\theta, 1/\psi_2)^{I(w_i\geq 0)}$$

$$\propto \prod_{i:w_i\geq 0} f_N(y_i|X_i\theta, 1/\psi_2)$$

If we define:

Σ_+ is the $n_2 * n_2$ diagonal matrix with the diagonal elements are $1/\psi_2$'s.

X_+ is the submatrix of X such columns i that $w_i \geq 0$

y_+ is the subvector of y such elements i that $w_i \geq 0$, then:

$$f(\theta|...) \propto |\Sigma_+|^{-\frac{1}{2}} exp(-\frac{1}{2}(y_+ - X_+\theta)^T\Sigma_+^{-1}(y_+ - X_+\theta))$$

$$\sim N((X_+^T\Sigma_+^{-1}X_+)^{-1}X_+^T\Sigma_+^{-1}y_+, (X_+^T\Sigma_+^{-1}X_+)^{-1})$$

This is the bivariate normal with mean $(X_+^T\Sigma^{-1}X_+)^{-1}X_+^T\Sigma^{-1}y_+$ and covariance matrix $(X_+^T\Sigma^{-1}X_+)^{-1}$.

2.2.2 Model 2

Compared to model 1, this model has one fewer parameter and the full conditional distributions have been changed accordingly.

1 The full conditional for w:

$$f(w|...) \propto \prod_{i=1}^{n}(f_N(y_i|\mu_1, 1/\psi_1)I(w_i < 0) + f_N(y_i|\theta_s * \mu_1 + \theta * x_i, 1/\psi_2)I(w_i \geq 0))f_N(w_i|X_i\beta, 1)$$

2 The full conditional for μ_1:

$$f(\mu_1|...) \propto \prod_{i=1}^{n} f_N(y_i|\mu_1, 1/\psi_1)I(w_i < 0) + f_N(y_i|\theta_s * \mu_1 + \theta * x_i, 1/\psi_2)$$
$$\sim N(\frac{\psi_1 \sum_{i:w_i<0} y_i + \theta_s\psi_2 \sum_{i:w_i\geq 0}(y_i - \theta * x_i)}{n_1\psi_1 + \theta_s n_2\psi_2}, \frac{1}{n_1\psi_1 + \theta_s n_2\psi_2})$$

This is still a univariate normal distribution but the parameters are different compared with model 1.

3 The full conditional for ψ_1:

$$f(\psi_1|...) \propto \psi_1^{-1} \prod_{i=1}^{n} f_N(y_i|\mu_1, 1/\psi_1)^{I(w_i<0)} f_N(y_i|\theta_s * \mu_1 + \theta * x_i, 1/\psi_2)^{I(w_i\geq 0)}$$
$$\sim \psi_1^{\frac{n_1}{2}-1} exp(-\frac{1}{2}\psi_1 \sum_{i:w_i<0} (y_i - \mu_1)^2)$$

4 The full conditional for ψ_2:

$$f(\psi_2|...) \propto \psi_2^{-1} \prod_{i=1}^{n} f_N(y_i|\mu_1, 1/\psi_1)^{I(w_i<0)} f_N(y_i|\theta_s * \mu_1 + \theta * x_i, 1/\psi_2)^{I(w_i\geq 0)}$$
$$\sim \psi_2^{\frac{n_2}{2}-1} exp(-\frac{1}{2}\psi_2 \sum_{i:w_i\geq 0} (y_i - \theta_s * \mu_1 - \theta * x_i)^2)$$

5 The full conditional for β:

$$f(\beta|...) \propto \prod_{i=1}^{n} f(w_i|X_i, \beta)$$
$$\propto exp(-\frac{\sum_{i=1}^{n}(w_i - X_i\beta)^2}{2})$$

This is the bivariate normal distribution with mean $(X^TX)^{-1}X^Tw$ and covariance matrix $(X^TX)^{-1}$.

6 The full conditional for θ:

$$f(\theta|...) \propto \prod_{i=1}^{n} f_N(y_i|\mu_1, 1/\psi_1)I(w_i < 0) + f_N(y_i|\theta_s * \mu_1 + \theta * x_i, 1/\psi_2)$$

$$\propto \prod_{i:w_i \geq 0} f_N(y_i|\theta_s * \mu_1 + \theta * x_i, 1/\psi_2)$$

$$\sim N(\frac{\sum_{i:w_i \geq 0}(y_i - \theta_s * \mu_1)x_i}{\sum_{i:w_i \geq 0} x_i^2}, \frac{1}{\psi_2 \sum_{i:w_i \geq 0} x_i^2})$$

2.3 Simulation Study

2.3.1 Data Generation

We conduct a simulation study to examine the proposed models based on the data set collected on Interstate I-35 near San Antonio, Texas (Guo et al. 2010). The study corridor covered a sixteen kilometer section with an average daily traffic volume around 150,000 vehicles. The travel time was collected when vehicles tagged using a radio frequency device passed the automatic vehicle identification stations on New Braunfels Ave. (Station no. 42) and OConnor Rd. (Station no. 49). The Figure 2.1 illustrates the data collection procedures.

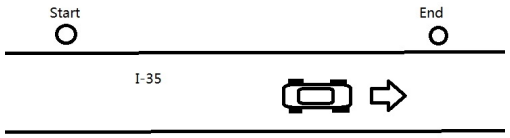

Figure 2.1: Illustration of Data Collection

Based on the definition, the traffic volume is the number of vehicles travel through the road segment during one specific hour period (e.g. 7:00-7:59). The data set contains 237 distinct hours of observations. We average the traffic volume by the hour of a day. The Figure 2.2 illustrates that the range of average traffic volume by hour, which is roughly from 1 to 25.

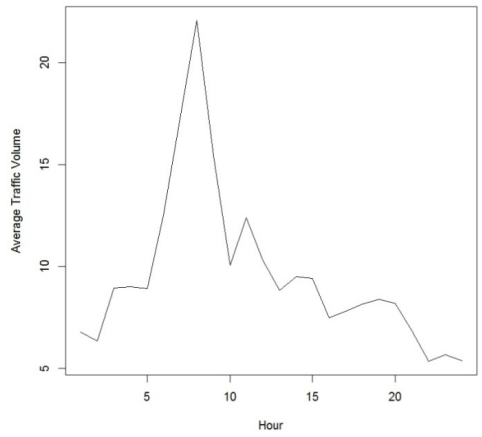

Figure 2.2: Average Traffic Volume by Hour of a Day

However, in the original data only a proportion of vehicles on th road (which were equipped with electronical identifications) were counted. In order to estimate the true traffic volume, we simulated new data sets according to the shape of the original data and extended it by a scale:

Y_{ij} : Simulated traffic volume of hour i in day j. $Y_{ij} = [c * \mu_i + \epsilon_{ij}]^+, \epsilon_{ij} \sim N(0, d^2)$

X_{ij} : Original traffic volume of hour i in day j. i=0...23, j= 1...10 or 11

μ_i : Average traffic volume of hour i. $\mu_i = \dfrac{\sum_k X_{ik}}{\text{Numbe of days}}$

Based on historical data and without loss of generality, we select d=100 and c=50. By above procedure, we generated the traffic volume and the travel time data sets accordingly.

2.3.2 Simulation Procedures

For a given model and a set of predetermined parameters, the simulation study is proceeded as follows:

```
1. Set n=Number of simulations we plan to run.
2. For (i in 1:n){
    Generate data
    Do{
          Markov Chain Monte Carlo
    }While convergence
    Record if the 95% credible intervals cover the true values
    }
```

For the ith round of MCMC procedure, we ran more than 5000 iterations in each MCMC procedure to ensure convergence. Finally, we could verify whether the frequencies that the credible intervals covered the true values were close to 95%. In the following sections, the simulation results will be discussed thoroughly.

2.3.3 Model 1 VS Model 2

The basic difference between model 1 and 2 is that in model 2 we have an extra parameter θ_s which could be determined via engineering expertise. θ_s controls the difference between free-flow mean travel time μ_1 and the baseline of congested mean travel time μ_2. It is tempting to set $\theta_s = 1$ but the initial analysis indicated that when θ_s is too close to 1 the identifiability issue will affect the model fitting. Therefore, we start our simulation study at $\theta_s = 1.2$.

The values of μ_1, ψ_1, ψ_2, β_0 and β_1 are set according to historical data. We have plot the relationship between probability in congested state and the traffic volume as Figure 2.3.

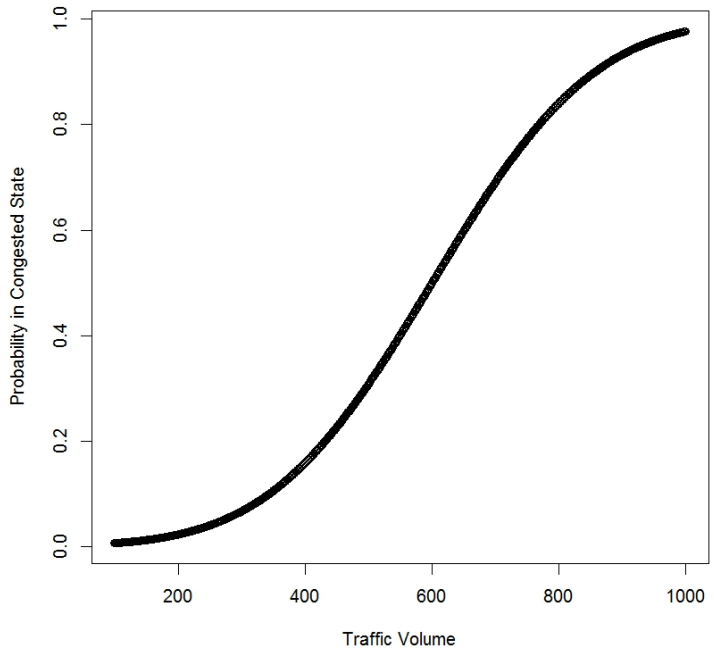

Figure 2.3: Relationship of Probability in Congested State and Traffic Volume

We have tried five different settings of $\theta_0(\theta_s)$ and θ_1. The results are summarized in following tables.

Table 2.2: Model 1 & 2: Average of Posterior Means Comparison

	μ_1	ψ_1	ψ_2	β_0	β_1	$\theta_0(\theta_s)$	θ_1
Setting 1	500	0.01	1.0e-4	-3	0.005	600 (1.2)	0.6
Model 1	499.9	0.010	9.8e-5	-2.93	0.0049	592.9	0.61
Model 2	499.9	0.010	9.8e-5	-2.94	0.0049		0.60
Setting 2	500	0.01	1.0e-4	-3	0.005	650 (1.3)	0.6
Model 1	499.9	0.010	9.3e-5	-2.98	0.0049	648.1	0.60
Model 2	500.0	0.010	9.9e-5	-2.98	0.0050		0.60
Setting 3	500	0.01	1.0e-4	-3	0.005	650 (1.3)	0.3
Model 1	499.9	0.0093	9.3e-5	-2.83	0.0047	628.2	0.32
Model 2	500.0	0.010	9.5e-5	-2.88	0.0048		0.30
Setting 4	500	0.01	1.0e-4	-3	0.005	700 (1.4)	0.3
Model 1	499.9	0.010	9.8e-5	-2.96	0.0049	694.8	0.31
Model 2	500.0	0.010	9.8e-5	-2.96	0.0049		0.30
Setting 5	500	0.01	1.0e-4	-3	0.005	750 (1.5)	0.3
Model 1	500.0	0.010	9.9e-5	-2.98	0.0049	747.1	0.30
Model 2	500.0	0.010	9.9e-5	-3.00	0.0050		0.30

Table 2.3: Model 1 & 2: Coverage Probabilities Comparison

	μ_1	ψ_1	ψ_2	β_0	β_1	$\theta_0(\theta_s)$	θ_1
Setting 1	500	0.01	1.0e-4	-3	0.005	600 (1.2)	0.6
Model 1	0.94	0.87	0.85	0.71	0.77	0.64	0.72
Model 2	0.93	0.91	0.79	0.69	0.78		0.97
Setting 2	500	0.01	1.0e-4	-3	0.005	650 (1.3)	0.6
Model 1	0.96	0.95	0.93	0.93	0.91	0.91	0.93
Model 2	0.97	0.96	0.95	0.95	0.93		0.97
Setting 3	500	0.01	1.0e-4	-3	0.005	650 (1.3)	0.3
Model 1	0.97	0.57	0.23	0.17	0.24	0.11	0.17
Model 2	0.95	0.76	0.36	0.36	0.49		0.90
Setting 4	500	0.01	1.0e-4	-3	0.005	700 (1.4)	0.3
Model 1	0.93	0.84	0.92	0.85	0.89	0.86	0.86
Model 2	0.96	0.90	0.89	0.86	0.90		0.95
Setting 5	500	0.01	1.0e-4	-3	0.005	750 (1.5)	0.3
Model 1	0.94	0.97	0.97	0.91	0.92	0.88	0.93
Model 2	0.98	0.99	0.92	0.96	0.96		0.93

Table 2.4: Model 2 Between Setting 2 and 3: Coverage Probabilities Comparison

Value of θ_1	μ_1	ψ_1	ψ_2	β_0	β_1	θ_1
0.3	0.95	0.76	0.36	0.36	0.49	0.90
0.4	0.89	0.87	0.75	0.67	0.74	0.93
0.5	0.94	0.92	0.84	0.90	0.91	0.94
0.6	0.97	0.96	0.95	0.95	0.93	0.97

The Table 2.2 shows that the point estimates of the parameters are generally good and Model 2 seems to be slightly better than Model 1 but the difference is minimal.

However, when the coverage probability is taken into account, the Table 2.3 shows that the estimates of β_0, β_1 and θ_1 could be quite off the target if the two components are close to each other. Under that situation, the Model 2 generaly outperforms Model 1, which is reasonable since we have got extra information (θ_s).

The Table 2.4 is a further investiagtion for Model 2 between Setting 2 and 3. Since the value of θ_s plays an important role in the coverage probability, we selected two more points between 03 to 0.6 and the coverage probabilities are shown. In general, when θ_s increases, the coverage probabilities are closer to the target 95%.

The Figure 2.4 is a comparison of coverage probabilities in Model 1 and 2 among 5 settings. The dashed line is used to denote 95% for reference. Overally Model 2 performs better than Model 1.

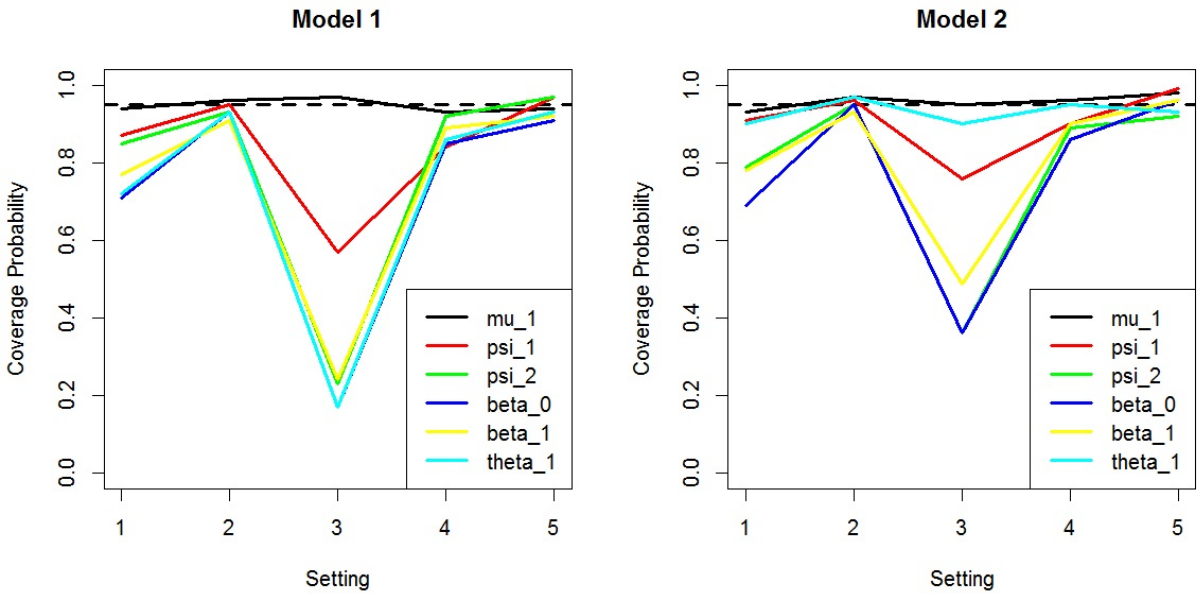

Figure 2.4: Model 1 VS Model 2: Coverage Probabilities Comparison in 5 Settings

Later we will show that if the θ_s is misspecified in the Model 2 the coverage probability of β_0 and β_1 could be even better at the cost of θ_1, which might be interesting.

2.3.4 Simulation Results of Model 2

For Model 2, we design a more complicated simulation study based on a set of combinations of different parameters. Some of the parameters are fixed among all the models:

$$\mu_1 = 500, \psi_1 = 0.01, \beta_0 = -3, \psi_2 = 0.0001$$

On the other hand, multiple levels of these parameters are applied:

$$\beta_1 \in \{0.004, 0.0045, 0.005\}$$

$$\theta \in \{0.3, 0.6\}$$

$$\theta_s \in \{1.2, 1.3, 1.4, 1.5\}$$

β_1 represents the relationship between proportion of congested state and the traffic volume. θ indicates the relationship between travel time under congested state and the traffic volume

There are 24 different plans for simulation. The output has been summarized as Table 2.5, which shows that the coverage probabilities are generally good when the two components are well separated and vice versa.

Table 2.5: More Results of Model 2: Coverage Probabilities

ID	β_1	θ	θ_s	Cov of μ_1	Cov of ψ_1	Cov of ψ_2	Cov of β_0	Cov of β_1	Cov of θ
1	0.004	0.3	1.2	0.78	0.09	0	0	0	0.23
2	0.004	0.3	1.3	0.89	0.71	0.09	0.09	0.24	0.78
3	0.004	0.3	1.4	0.91	0.94	0.61	0.84	0.88	0.92
4	0.004	0.3	1.5	0.95	0.95	0.93	0.96	0.96	0.94
5	0.004	0.6	1.2	0.91	0.87	0.71	0.69	0.71	0.96
6	0.004	0.6	1.3	0.95	0.94	0.91	0.89	0.89	0.96
7	0.004	0.6	1.4	0.95	0.93	0.93	0.93	0.94	0.93
8	0.004	0.6	1.5	0.93	0.98	0.96	0.94	0.97	0.93

Continued on next page

Table 2.5 – *Continued from previous page*

ID	β_1	θ	θ_s	Cov of μ_1	Cov of ψ_1	Cov of ψ_2	Cov of β_0	Cov of β_1	Cov of θ
9	0.0045	0.3	1.2	0.79	0.18	0	0	0.01	0.41
10	0.0045	0.3	1.3	0.91	0.77	0.24	0.19	0.29	0.84
11	0.0045	0.3	1.4	0.93	0.94	0.92	0.93	0.93	0.74
12	0.0045	0.3	1.5	0.94	0.92	0.93	0.91	0.94	0.92
13	0.0045	0.6	1.2	0.93	0.89	0.78	0.74	0.75	0.94
14	0.0045	0.6	1.3	0.97	0.99	0.94	0.93	0.93	0.96
15	0.0045	0.6	1.4	0.98	0.96	0.89	0.88	0.89	0.98
16	0.0045	0.6	1.5	0.97	0.94	0.95	0.96	0.94	0.89
17	0.005	0.3	1.2	0.83	0.26	0.1	0	0	0.66
18	0.005	0.3	1.3	0.95	0.76	0.36	0.36	0.49	0.9
19	0.005	0.3	1.4	0.96	0.9	0.89	0.86	0.9	0.95
20	0.005	0.3	1.5	0.98	0.99	0.92	0.96	0.96	0.93
21	0.005	0.6	1.2	0.93	0.91	0.79	0.69	0.78	0.97
22	0.005	0.6	1.3	0.97	0.96	0.95	0.95	0.93	0.97
23	0.005	0.6	1.4	0.92	0.93	0.95	0.92	0.93	0.93
24	0.005	0.6	1.5	0.94	0.94	0.94	0.95	0.96	0.87

The Figure 2.5 might be more intuitive than Table 2.5. The dashed line is used to denote

95% for reference. The larger θ and θ_s are, the higher the coverage probabilities will be. However, the relationship between coverage probability and β_1 is not direct.

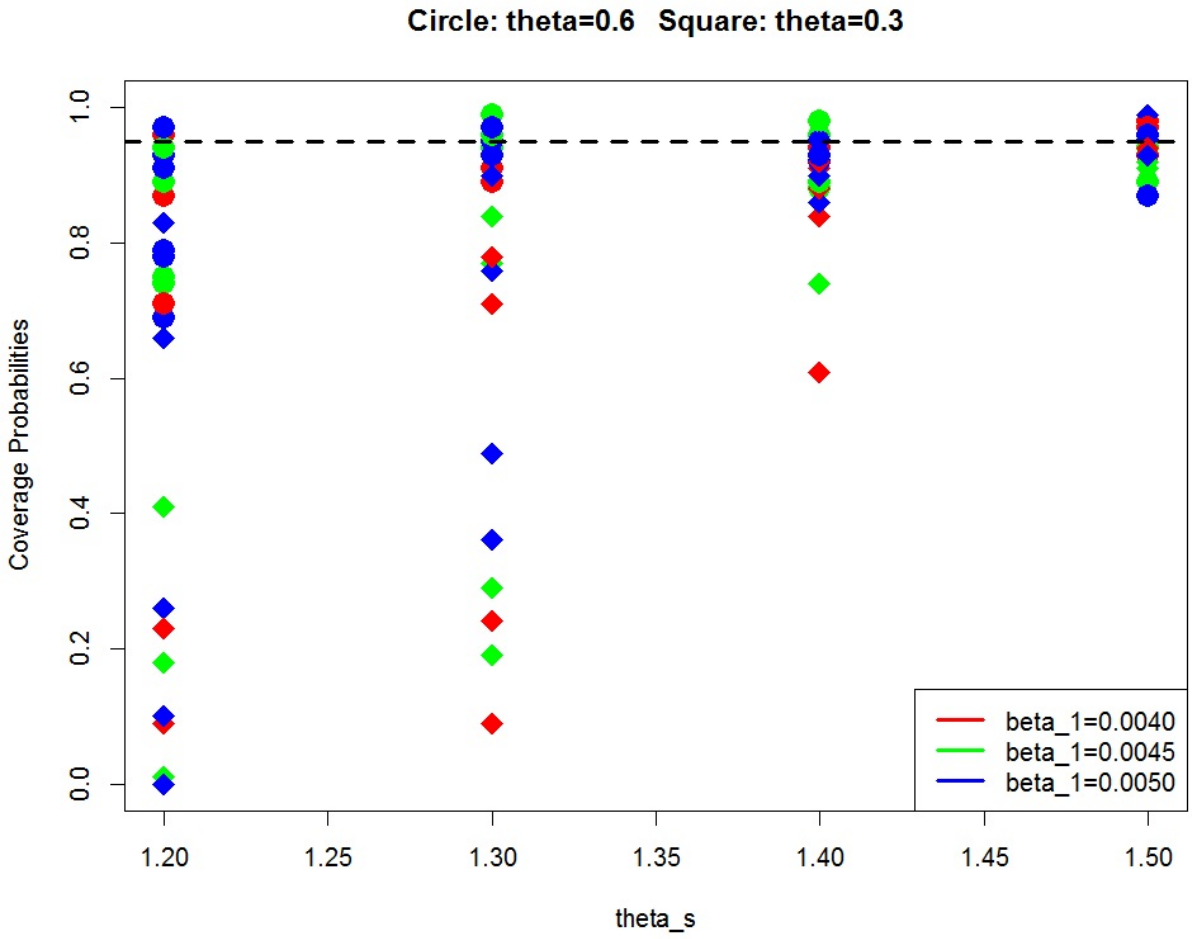

Figure 2.5: Model 2: Coverage Probabilities Comparison

2.3.5 Robustness of Misspecified θ_s

One of the basic problem of Model 2 is the choice θ_s. If the true value of θ_s is unknown and we misspecify it, it is useful to know whether the model is still able to provide meaningful result.

For example, what happens if we ran the MCMC based on $\theta_s = 1.3$ but it is actually 1.2? Therefore, we tried four different settings to verify the robustness of misspecified θ_s.

Table 2.6: Misspecified Models: Average of Posterior Means Comparison

	μ_1	ψ_1	ψ_2	β_0	β_1	θ_s	θ_1	μ_2
Setting 1	500	0.01	1.0e-4	-3	0.005	1.2	0.3	769
True Model	499.9	0.010	9.29e-5	-2.72	0.0046	1.2	0.30	774
Misspecified 1	499.9	0.010	9.26e-5	-2.84	0.0048	1.3	0.24	782
Misspecified 2	499.9	0.010	8.89e-5	-2.92	0.0049	1.4	0.18	800
Setting 2	500	0.01	1.0e-4	-3	0.005	1.2	0.6	939
True Model	500.0	0.010	9.81e-5	-2.94	0.0049	1.2	0.60	947
Misspecified	500.0	0.010	9.63e-5	-2.96	0.0050	1.3	0.54	958
Setting 3	500	0.01	1.0e-4	-3	0.005	1.3	0.3	819
True Model	500.0	0.010	9.52e-5	-2.87	0.0048	1.3	0.30	828
Misspecified 1	500.0	0.010	9.52e-5	-2.95	0.0048	1.4	0.24	837
Misspecified 2	500.0	0.010	9.03e-5	-2.98	0.0050	1.5	0.18	852
Setting 4	500	0.01	1.0e-4	-3	0.005	1.3	0.6	989
True Model	500.0	0.010	9.91e-5	-2.98	0.0050	1.3	0.60	989
Misspecified	500.0	0.010	9.74e-5	-2.99	0.0050	1.4	0.54	1005

Note: The μ_2 is estimated by $\theta_s * \mu_1 + \theta_1 * \bar{X}$

Table 2.7: Misspecified Models: Coverage Probabilities Comparison

	μ_1	ψ_1	ψ_2	β_0	β_1	θ_s	θ_1
Setting 1	500	0.01	1.0e-4	-3	0.005	1.2	0.3
True Model	0.91	0.30	0.09	0	0	1.2	0.71
Misspecified 1	0.78	0.67	0.09	0.18	0.28	1.3	0.28
Misspecified 2	0.78	0.81	0	0.61	0.67	1.4	0
Setting 2	500	0.01	1.0e-4	-3	0.005	1.2	0.6
True Model	0.96	0.94	0.78	0.73	0.77	1.2	0.95
Misspecified	0.90	0.95	0.59	0.91	0.95	1.3	0
Setting 3	500	0.01	1.0e-4	-3	0.005	1.3	0.3
True Model	0.95	0.73	0.42	0.35	0.42	1.3	0.88
Misspecified 1	0.90	0.90	0.41	0.84	0.85	1.4	0
Misspecified 2	0.89	0.95	0.02	0.88	0.92	1.5	0
Setting 4	500	0.01	1.0e-4	-3	0.005	1.3	0.6
True Model	0.94	0.88	0.89	0.94	0.95	1.3	0.93
Misspecified	0.92	0.92	0.71	0.92	0.91	1.4	0

Table 2.6 indicates that the point estimates of the parameters are generally stable, while the misspecified models generally push the mean of component 2 to the right. From Table 2.7, it can be concluded that when the two components are not well separated (i.e. θ_s and θ are small), the misspecified models can sometimes have better coverage of the regression

coefficients for proportion (ψ_1, β_0 and β_1).

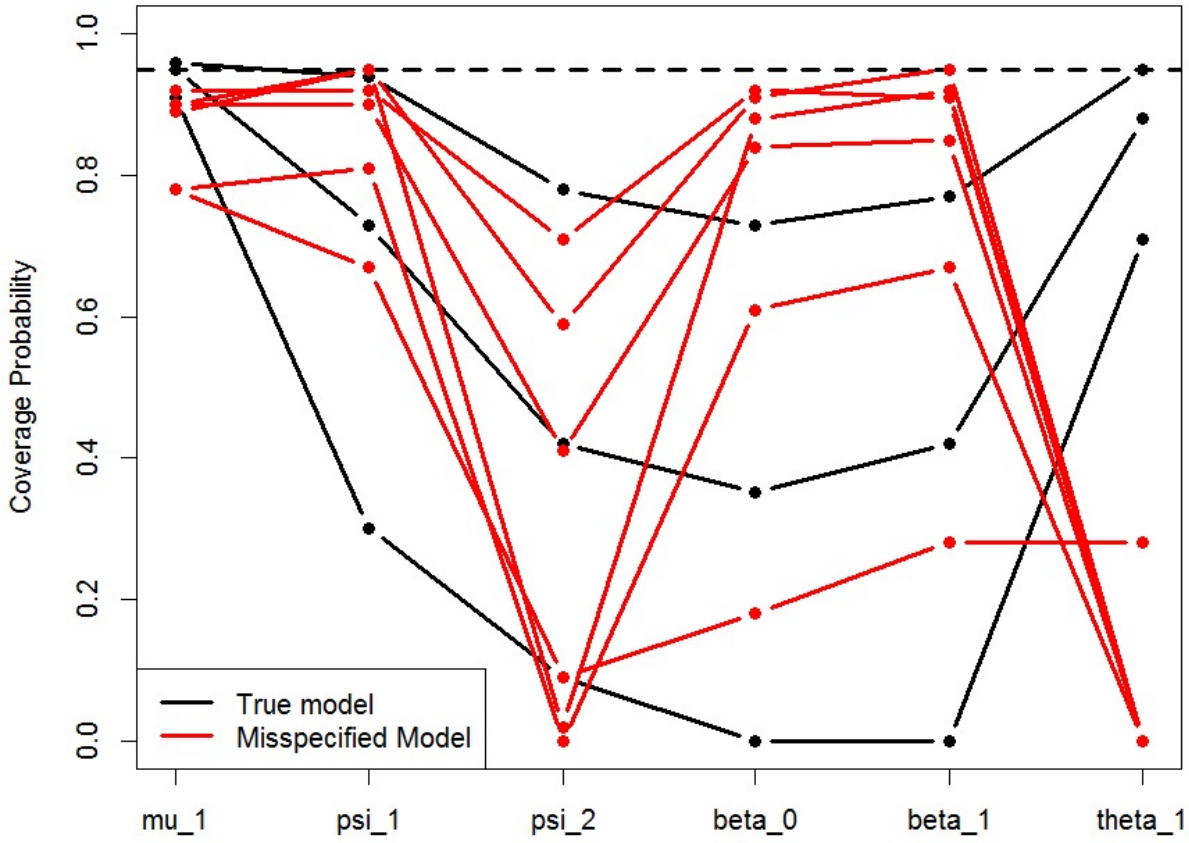

Figure 2.6: Misspecified and True Model Comparison

Figure 2.6 shows the coverage probabilities comparison between true and misspecified models in 4 different settings. The dashed line is used to denote 95% for reference. It seems that the true models are superior in estimating θ_1 and ψ_2, while misspecified models perform better in ψ_1, β_0 and β_1.

Although misspecified models are generally not good at estimating θ_1, researchers usually concern $\mu_2 = \theta_s * \theta_0 + \theta_1 * x$ rather than θ_1 itself. In order to evaluate the influence of misspecified θ_s on μ_2, we plot the relationship between traffic volume and the corresponding μ_2 under theoretical result, true model estimate and misspecified model estimate. Figure 2.7 shows that the misspecified model estimates are quite close to the theoretical results when the traffic volume is high. Actually, this is the case that the congested state is defined. Therefore, the application of these models are still reasonable when θ_s is misspecified.

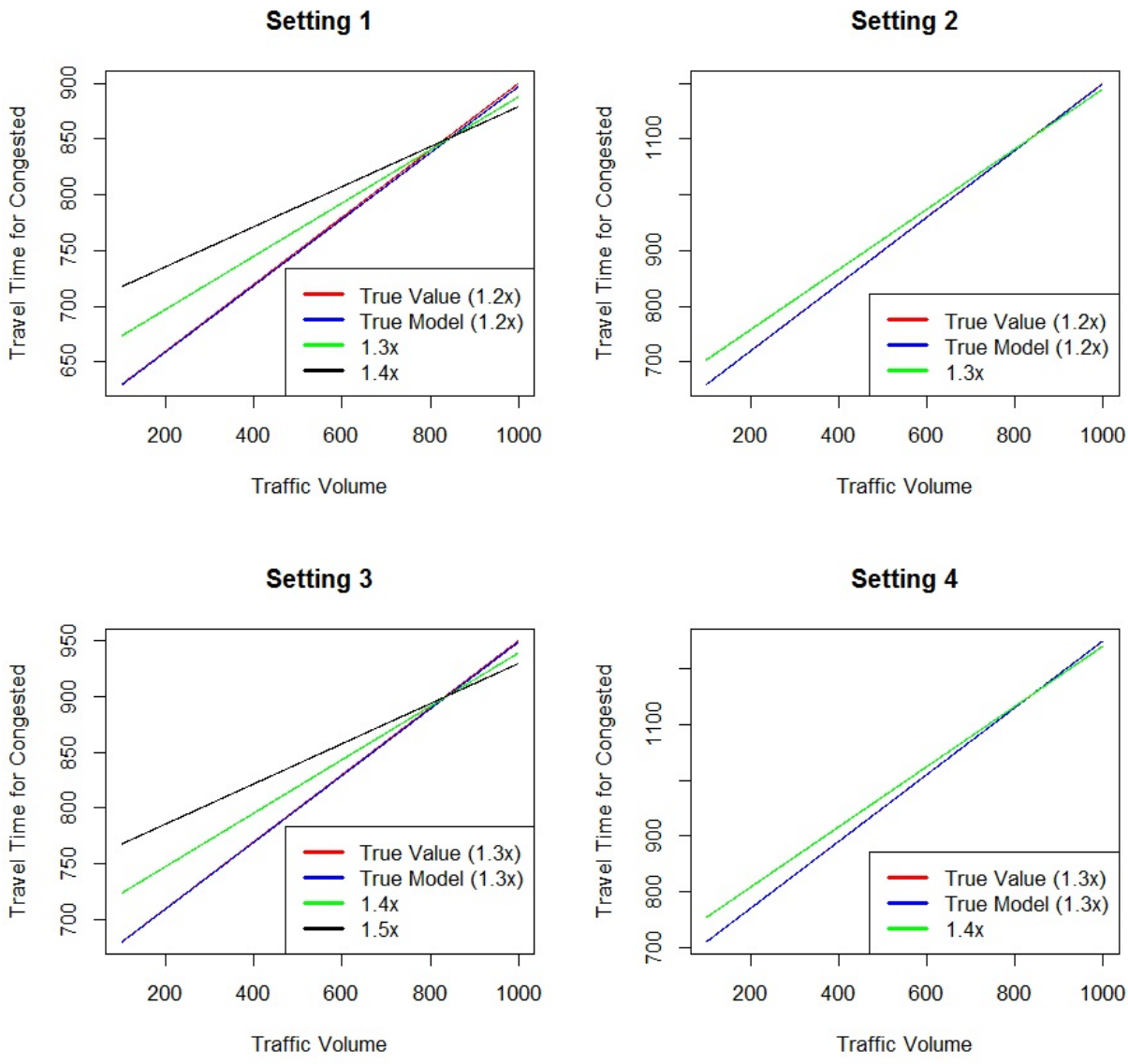

Figure 2.7: Theoretical, Misspecified and True Model Comparison

2.4 Model with Real Data

We apply Model 2 to the real data and choose different values of θ_s. The initial results are as follows:

Table 2.8: Results from Real Data with Different $\theta_s's$

Parameter	$\theta_s = 1$	$\theta_s = 1.1$	$\theta_s = 1.2$	$\theta_s = 1.3$	$\theta_s = 1.4$
μ_1	578.6	578.5	578.3	578.3	578.2
ψ_1	0.00083	0.00083	0.00083	0.00082	0.00081
ψ_2	1.01e-5	9.99e-6	9.68e-6	9.29e-6	8.86e-6
β_0	-0.97	-1.00	-1.04	-1.09	-1.13
β_1	0.031	0.033	0.036	0.038	0.040
θ	15.90	12.25	8.68	5.19	1.76
μ_2	758.6	774.7	792.0	810.3	829.4

Note: The μ_2 is estimated by $\theta_s * \mu_1 + \theta * \bar{X}$

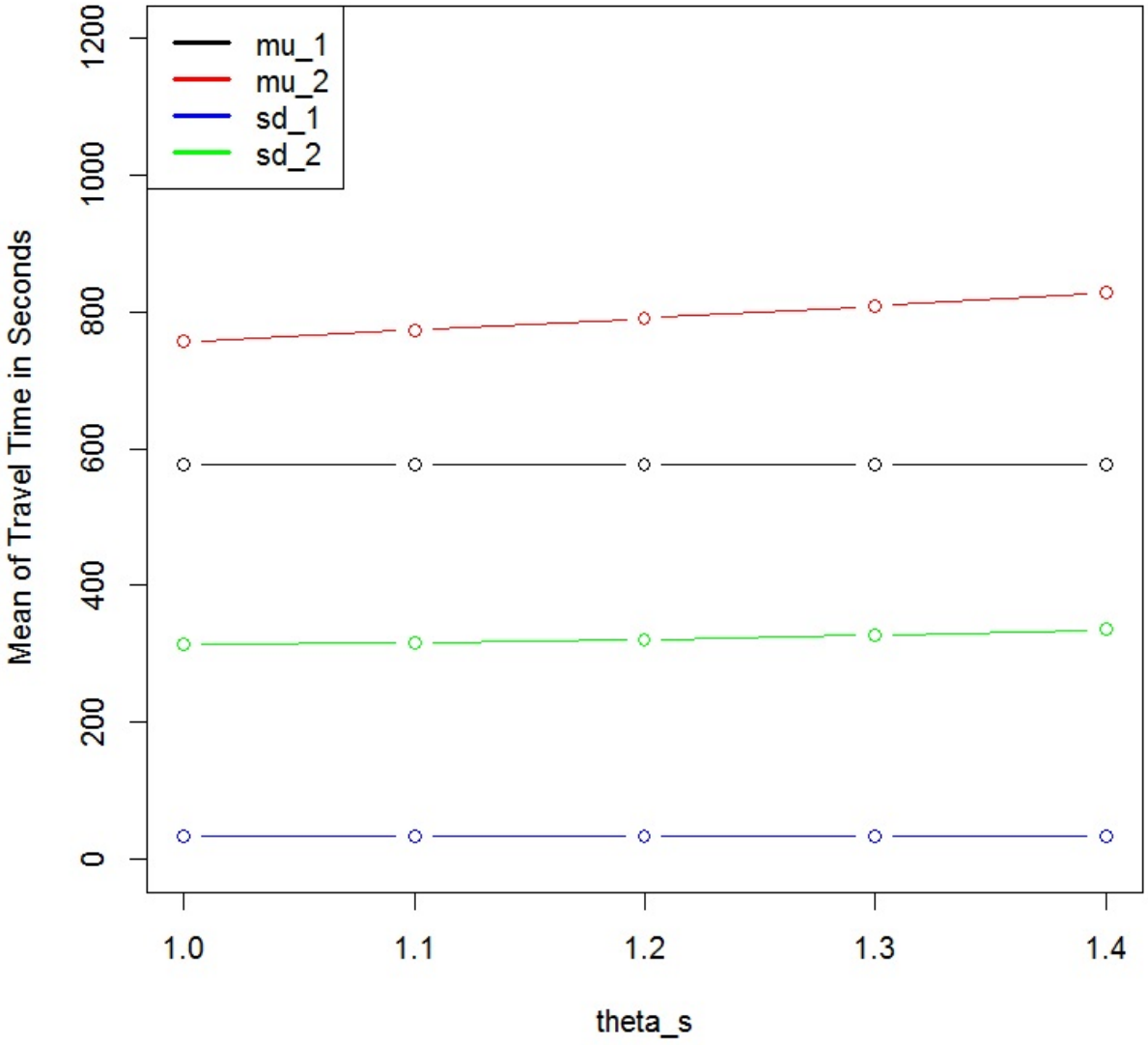

Figure 2.8: Parameters Estimates under Different θ_s'

Figure 2.8 shows the relationship between some of the important parameters estimates and θ_s. Both the means and the standard deviations of the two componenets are quite stable

with respect to the change of θ_s.

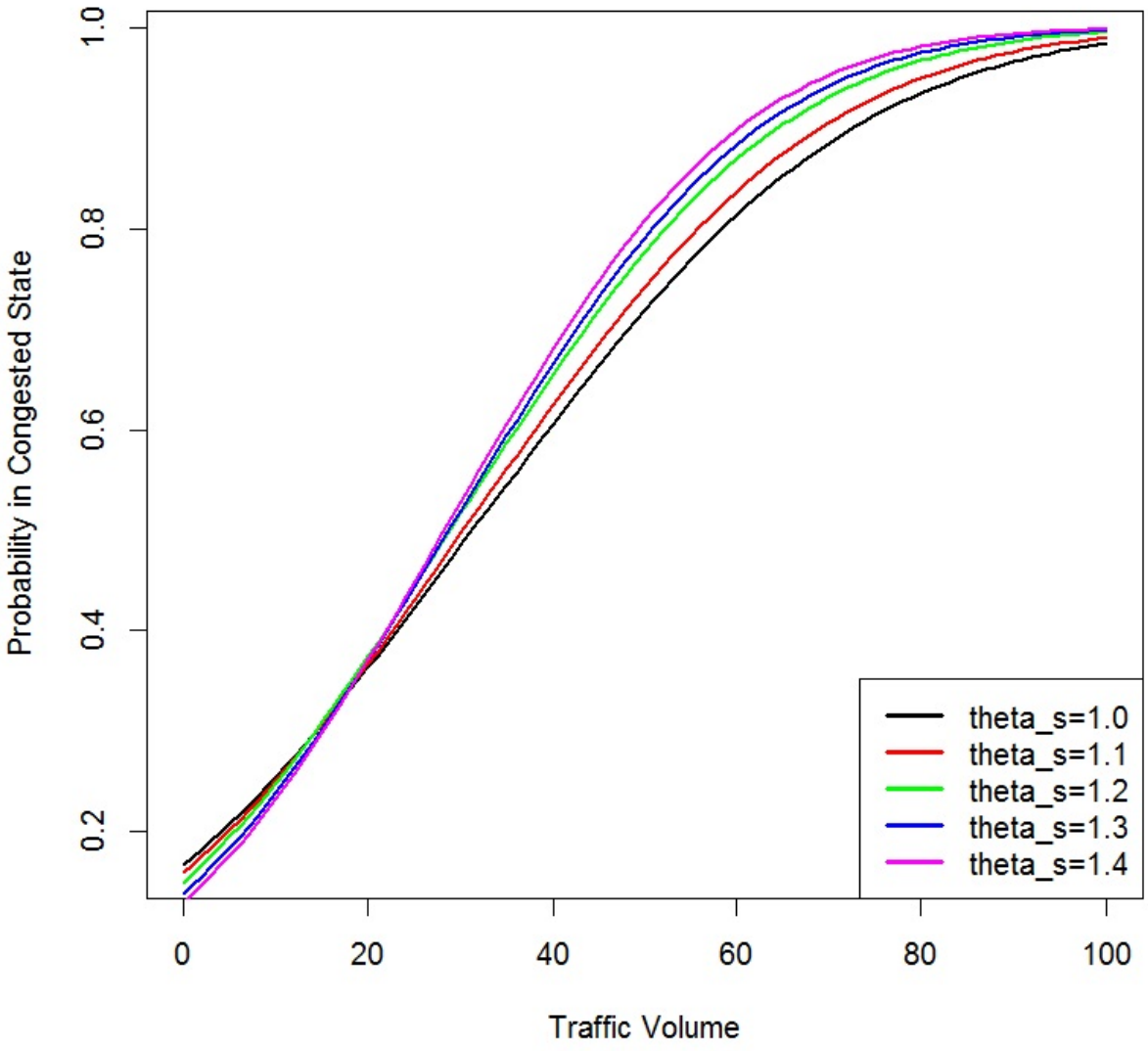

Figure 2.9: Probability in Congested State and Traffic Volume: Real Data

Figure 2.9 indicates the relationship between probability in congested State and traffic volume under different settings of θ_s.

2.5 Summary

The multi-state model provides a flexible and efficient framework for modeling travel time reliability, especially under complex traffic conditions. Guo et al. (2012) illustrated that the multi-state model outperforms single-state models in congested or near-congested traffic conditions and the advantage is substantial in high traffic volume condition.

Our objective is to quantitatively evaluate the influence of traffic volume on the mixture of two components. Our work advanced the multi-state models by proposing regressions on the proportions and distribution parameters for underlying traffic states. The Bayesian analysis also provides feasible credible intervals for each parameters without asymptotic assumption.

Previous studies usually model the travel time solely without establishing the relationship between travel time and important transportation statistics such as traffic volume. Our model can also be easily extended to include more covariates in either linear or nonlinear form.

Our modeling result indicates that there is a negatively relationship between the proportion of free-flow state and the traffic volume, which confirms the statement raised by Guo et al. (2012) that for low traffic volume condition, there might only exist one travel time state and single-state models will be sufficient. The estimation for the congested state indicates that the travel time under such condition exhibits substantial variability and positively related with traffic volume, which also verifies the phenomenon found by Guo et al. (2012).

There are several potential extensions to the current research. Current research only includes lognormal and normal distributions. A number of other distributions, e.g., Gamma and extreme value distributions can also be investigated. Finally, one of the assumption for the existing Bayesian mixture model is that all the observations are independent, which could be relaxed in the Hidden Markov model. The Hidden Markov model will be the topic of the next chapter.

Chapter 3

Travel Time Reliability: Hidden Markov Model

3.1 Introduction

The Bayesian mixture regression model discussed in previous Chapter is based on the assumption that the observations are independent. In I-35 data the travel time of vehicles were collected chronologically, which implied that the independence assumption might be unrealistic. It is possible to apply auto-correlated error terms to handle this problem (Cochrane and Orcutt 1949), but the interpretation regarding this scenario is unclear.

In order to accommodate the dependency structure of the data, we decided to follow a gentle method: Hidden Markov model. The basic concept of hidden Markov model was introduced

by Baum and Petrie (1966). It also includes traditional mixture model as a special case (Scott 2002).

Hidden Markov models have become important in a wide variety of applications including (Couvreur 1996): speech recognition (Rabiner 1989), biometrics (Albert 1991), econometrics (Hamilton 1989), computational biology (Krogh et al. 1994), fault detection (Smyth 1994) and many other areas.

3.2 Autocorrelation

In this section we will analyze the I-35 data as a time series and the autocorrelation will be calculated. Autocorrelation (Wiener 1930) is a measure of similarity between observations with certain time lags.

For a sequence $\{X_t\}$, the autocorrelation of time lag s is defined as:

$$ACF(t,s) = \frac{E(X_t - \mu_t)(X_s - \mu_s)}{\sigma_t \sigma_s}$$

If we assume that $\{X_t\}$ is second-order stationary (Wold 1938), the autocorrelation can be written as:

$$ACF(s) = \frac{E(X_t - \mu)(X_{t+s} - \mu)}{\sigma^2}$$

For independent sequence, it is easy to see that ACF(s) should be small no matter the value of s is. For a sequence as $\{X_t : X_t = 0.5 * X_{t-1} + \epsilon_t\}$, the ACF(s) will be quite large when s=1 and decreases gradually as s increases.

In Figure 3.1, two plots are shown and to be compared. The x-axis is the time lag, while the y-axis is the ACF. The plot on the left shows the ACF of independent sequence while the plot on the right is estimated from the I-35 data. It is clear that the observed travel time is not an independent data set.

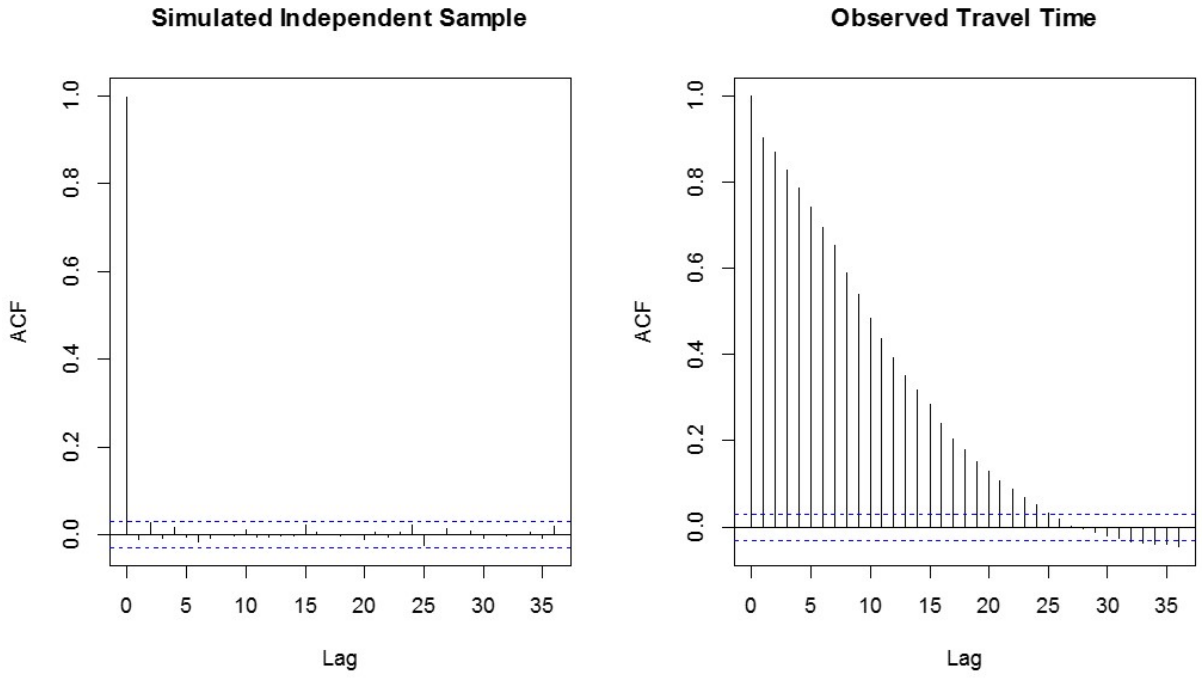

Figure 3.1: Autocorrelation Comparison

A formal test of the autocorrelation is Durbin Watson test (Durbin and Watson 1950). The DurbinWatson statistic is defined as:

$$d = \frac{\sum_{t=2}^{T}(X_t - X_{t-1})^2}{\sum_{t=1}^{T} X_t^2}$$

The value of d is always between 0 and 4. If the Durbin Watson statistic is substantially less than 2, there might be positive correlation. On the other hand, there might be negative

correlation. Under the normal assumption, the null distribution of the Durbin Watson statistic is a linear combination of chi-squared variables.

To satisfy the normal assumption, we use Box-Cox transformation (Choongrak 1959): $x \to \frac{x^{\lambda}-1}{\lambda}$. Figure 3.2 indicates that $\lambda = -4$ is the optimal choice.

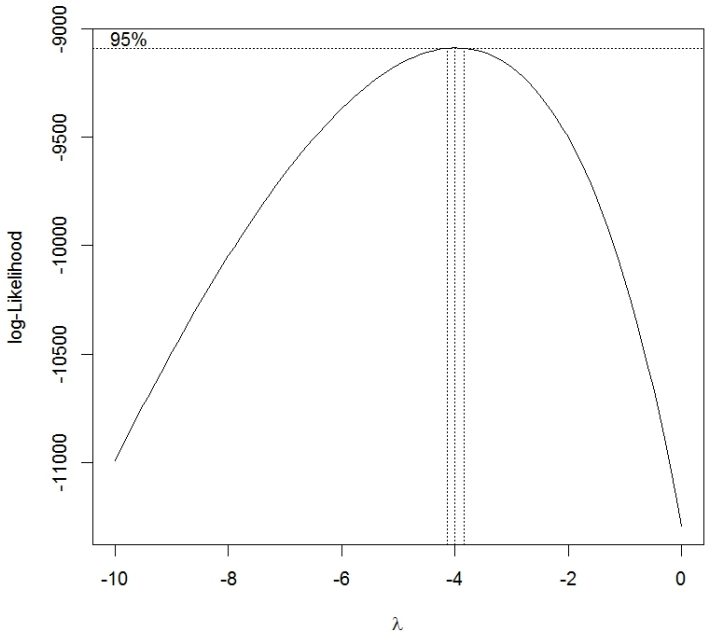

Figure 3.2: Box-Cox Transformation

The Durbin Watson statistic from the (transformed) travel time is 0.8244, which yields a p-value close to zero. This implies that the travel time data contains considerable positive autocorrelation.

3.3 Theoretical Background

3.3.1 Model Specification

A hidden Markov model consists of two sequences: the observed sequence $\{x_t\}, t = 1, 2, ..., n$ and the latent state sequence $\{s_t\}, t = 1, 2, ..., n$. Given the s_t, the distribution of observed data x_t is fully determined by the value of s_t. For example, if we denote the travel time in seconds as $\{x_t\}, t = 1, 2, ..., n$, then the sequence s_t could be defined as:

$$s_t = \begin{cases} 1 & \text{if the road is under free-flow} \\ 2 & \text{if the road is under congestion} \end{cases}$$

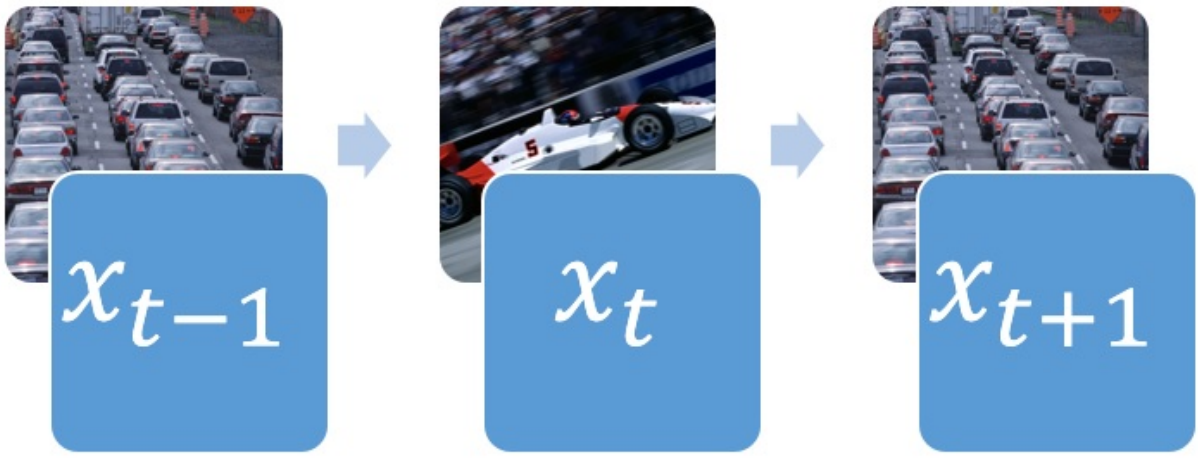

Figure 3.3: Hidden Markov Model: An Illustration

The two values of s_t represents the two states, which correspond to the two components in a mixture distribution. Given the state, the observed data x_t follows one of the distribu-

tions:

$$f(x_t|s_t) = \begin{cases} f(x|\Theta_1), & \text{if } s_t = 1 \\ f(x|\Theta_2), & \text{if } s_t = 2 \end{cases}$$

The form of the distribution $f(x|\Theta)$ could be normal, Gamma, Poisson, multinomial or others. For example, $x_t|s_t = 1 \sim N(1000, 100^2)$ and $x_t|s_t = 2 \sim N(500, 30^2)$.

As we mentioned before, the term "Hidden" indicates that $\{s_t\}$ is a latent sequence which can not be observed. Secondly, the term "Markov" indicates an important property of $\{s_t\}$:

$$P(s_t|s_{t-1}, ..., s_1) = P(s_t|s_{t-1}), \forall t \geq 2$$

Thus, $\{s_t\}$ is a Markov chain and has its transition probability matrix. For a two-state sequence, the transition matrix is as follow:

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix},$$

where P_{ij} is the probability that $P(s_{t+1} = j|s_t = i)$.

It is easy to see that, if $\{s_t\}$ is a trivial Markov Chain, i.e. i.i.d. then the hidden Markov model is equivalent to a traditional mixture model. Figure 3.4 is an illustration of two-state Markov chain.

Figure 3.4: Illustration of Two States Markov Chain

Here are some properties a Markov chain might have:

Irreducible: It is possible to get to any state from any state;

Aperiodic: A state i has period k ($k > 2$) if $\{n : p_{ii}^{(n)} > 0\} = \{k * d : d \geq 1\}$. If none of the state is periodic, the chain is aperiodic;

Positive recurrent: A state i is positive recurrent if the expected time that state i returns to itself is finite.

If every state in an irreducible chain is positive recurrent, there exists an unique stationary distribution π that satisfies:

$$\pi P = \pi$$

If an irreducible chain is positive recurrent and aperiodic, it is said to have a limiting distri-

bution ϕ:

$$\lim_{n\to+\infty} p_{ij}^{(n)} = \phi_j,$$

$$\text{where } p_{ij}^{(n)} = P(s_{i+n} = j | s_i = i)$$

A limiting distribution, when it exists, is always a stationary distribution, but the converse is not true. Stationary, or limiting distribution can be used to address the long-term behaviour of a Markov chain. For example, suppose a hidden Markov model has such transition matrix:

$$P = \begin{pmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{pmatrix},$$

By solving:

$$\begin{cases} \pi_1 = 0.8\pi_1 + 0.1\pi_2 \\ \pi_1 + \pi_2 = 1 \end{cases}$$

We conclude that $\pi_1 = 1/3, \pi_2 = 2/3$. Roughly speaking, 1/3 of the observations will be in state 1 while 2/3 of the observations will be in state 2 in a long-term run. This relates the hidden Markov model with the traditional mixture model.

One of our interests is to evaluate the influence on the travel time from traffic volume. Therefore, we want to conduct the regression models on transition probabilities using the traffic volume data. Hereafter we use y to denote the observed data and x as the covariate.

When the hidden Markov model has only two states, the transitiona matrix can be modeld in the style as logistic regression. In this case, the transition probability matrix is a 2×2 matrix. Due to the constraints that $P_{11} + P_{12} = P_{21} + P_{22} = 1$, the matrix has two free

parameters. Chung et al. (2007) discussed a similar model. For each row of the transition matrix, two logistic regression models with one covariate are proposed:

$$\log(\frac{P_{12}}{P_{11}}) = \beta_{0,1} + \beta_{1,1}x$$
$$\log(\frac{P_{22}}{P_{21}}) = \beta_{0,2} + \beta_{1,2}x$$

When the Markov chain has more than two states, multinomial logistic regression model will be applied. The first column are chosen as baseline. For example, the three states model is:

$$\log(\frac{P_{12}}{P_{11}}) = \beta_{0,1} + \beta_{1,1}x \quad \log(\frac{P_{13}}{P_{11}}) = \beta_{0,2} + \beta_{1,2}x$$
$$\log(\frac{P_{22}}{P_{21}}) = \beta_{0,3} + \beta_{1,3}x \quad \log(\frac{P_{23}}{P_{21}}) = \beta_{0,4} + \beta_{1,4}x$$
$$\log(\frac{P_{32}}{P_{31}}) = \beta_{0,5} + \beta_{1,5}x \quad \log(\frac{P_{33}}{P_{31}}) = \beta_{0,6} + \beta_{1,6}x$$

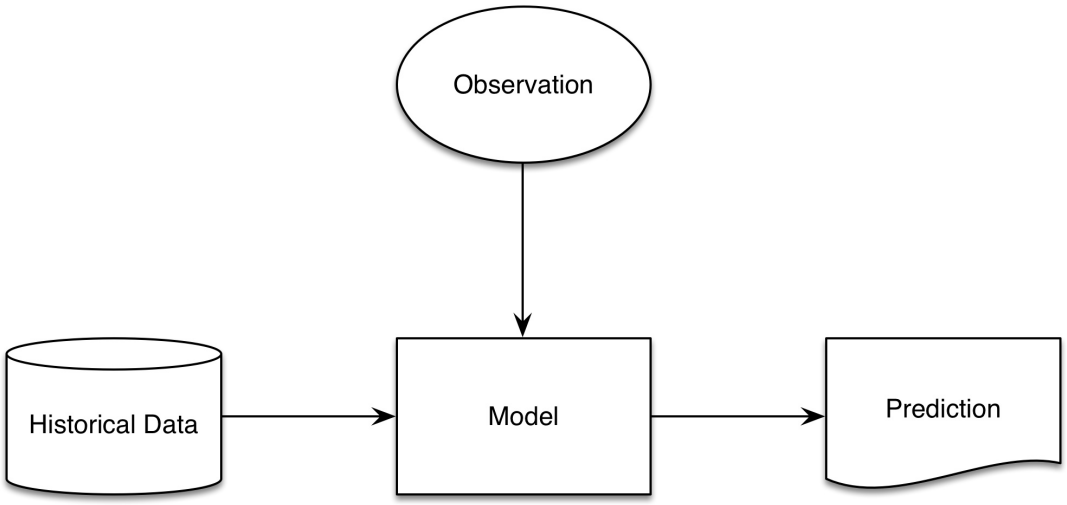

Figure 3.5: Hidden Markov Model: Flow Chart

Figure 3.5 illustrates the basic infrastructure of hidden Markov model. Both the historical data (traffic volume and travel time) and the observed data (real-time traffic volume) can be used to do prediction.

3.3.2 Model Estimation

There are several methods to estimate the parameters in the Hidden Markov model. One of the popular methods is EM algorithm (Baum et al. 1970, Bilmes 1998). Alternative ways include Viterbi and Gradient algorithm (Rabiner 1989). Bayesian method (Jean-Luc

and Chin-Hui 1991) has also been proposed. There are some existing software packages specifically for Hidden Markov model in R (Visser and Speekenbrink 2010). Although a Bayesian approach to HMM analysis does show some advantages in the more complex models, Rydn (2008) claimed that the results are generally similar and it is sufficient to use EM algorithm in most practical problems.

If we define:

$$L_k(t) = P(s_t = k|\mathbf{X})$$

$$H_{k,l}(t) = P(s_t = k, s_{t+1} = l|\mathbf{X})$$

The $L_k(t)$ is the conditional probability of being at state k at time t given the entire observed sequence $\mathbf{X}$. The $H_{k,l}(t)$ is the conditional probability of being at state k at time t and being at state l at time $t+1$ given the entire observed sequence $\mathbf{X}$.

The initial probabilities of state k (k=1,...,M) can be estimated by:

$$P(s_1 = k) \propto \sum_{t=1}^{T} L_k(t),$$

$$\sum_{k=1}^{M} P(s_1 = k) = 1$$

The EM algorithm can be described as follows ((Li and Gray 2000)):

- E step

  Compute $L_k(t)$ and $H_{k,l}(t)$ under current parameters values.

- M step

$$\mu_k = \frac{\sum_{t=1}^{T} L_k(t) x_t}{\sum_{t=1}^{T} L_k(t)}$$

$$\Sigma_k\text{: Covariance} = \frac{\sum_{t=1}^{T} L_k(t)(x_t - \mu_k)(x_t - \mu_k)^T}{\sum_{t=1}^{T} L_k(t)}$$

$$P(s_{t+1} = l | s_t = k)\text{: Transition probability} = \frac{\sum_{t=1}^{T-1} H_{k,l}(t)}{\sum_{t=1}^{T-1} L_k(t)}$$

In order to proceed E step, forward-backward algorithm can be used.

Define:

$$a_k(x_1, ..., x_t) = P(x_1, ..., x_t, s_t = k)$$

$$b_k(x_{t+1}, ..., x_T) = P(x_{t+1}, ..., x_T | s_t = k)$$

The forward algorithm is:

$$a_k(x_1) = P(s_1 = k) * f_k(x_1)$$

$$a_k(x_1, ..., x_t) = f_k(x_t) \sum_{i=1}^{M} a_i(x_1, ..., x_{t-1}) p_{ik}$$

f_k is the probability density of component k, p_{ik} is the transition probability.

The backward algorithm is:

$$b_k(x_{T+1}, ..., x_T) = 1 \text{ (Arbitrary setting)}$$

$$b_k(x_{t+1}, ..., x_T) = \sum_{i=1}^{M} p_{ki} f_i(x_{t+1}) b_i(x_{t+2}, ..., x_T)$$

Then $L_k(t)$ and $H_{k,l}(t)$ can be estimated as:

$$L_k(t) = \frac{a_k(x_1, ..., x_t) b_k(x_{T+1}, ..., x_T)}{\sum_{i=1}^{M} a_i(x_1, ..., x_t) b_i(x_{T+1}, ..., x_T)}$$

$$H_{k,l}(t) = \frac{a_k(x_1, ..., x_t) p_{kl} f_l(x_{t+1}) b_k(x_{t+1}, ..., x_T)}{\sum_{i=1}^{M} \sum_{j=1}^{M} a_i(x_1, ..., x_t) p_{ij} f_j(x_{t+1}) b_k(x_{t+1}, ..., x_T)}$$

The function $L_k(t)$ can be used to estimate the state an observation belongs to. However, the estimation is based on individual observation and may cause some problems. For example, it is possible that the result shows that $s_t = 1$ and $s_{t+1} = 2$. However, $p_{12} = 0$ which makes the entire sequence meaningless. Therefore, Viterbi algorithm (Viterbi 1967) might be applied to obtain the sequence with largest posterior probability.

3.3.3 Boostrap and Confidence Interval

In classical statistics, confidence interval plays an important role in model estimation. Visser et al. (2000) proposed several ways to obtain the confidence interval of hidden Markov models: finite approximation of Hessian, profile likelihood and bootstrap. He claimed that the results from first one are usually too narrow. Therefore, we will discuss about the other two methods here.

The profile likelihood method is based on profile likelihood ratio and χ^2 distribution. The basic idea is to evaluate the change of the log-likelihood caused by a single parameter when we treat all the other parameters as nuisance (Meeker and Escobar 1995). A profile likelihood for parameter β is defined as the likelihood function that all the other parameters are fixed at their MLE's:

$$PL(\beta) = \max_{\delta} L(\beta; \delta)$$

Suppose the MLE of β is $\hat{\beta}$, it can be shown that:

$$-2 * (logPL(\beta) - logPL(\hat{\beta})) \sim \chi^2(1) \text{ asymptotically.}$$

Based on the $\chi^2(1)$ distribution, we may derive the lower and upper bounds of the confidence interval easily. Figure 3.6 is an intuitive illustration, where Bm is the MLE while Bu and Bl are the upper and lower bounds.

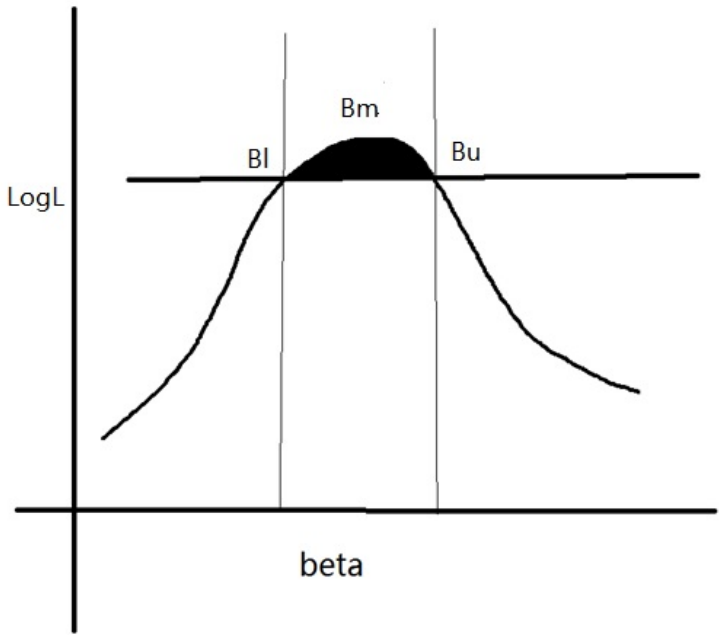

Figure 3.6: Confidence Interval by Profile Likelihood

The bootstrap idea is a popular technique to obtain confidence interval (Efron and Tibshirani 1994). However, a naive resampling method is not appropriate for hidden Markov model because that will destroy the dependency structure of the original data. Parametric bootstrap can be applied to handle this issue. There are generally three ways to do parametric bootstrap in hidden Markov model.

1 Based on parameter estimation.

2 Based on original data.

3 Mixture of 1 and 2.

For the first one, the paramours are estimated by original data and a new data set will be simulated solely based on the parameter estimates. The basic assumption for this method is that the model is correctly specified.

For the second one, after model fitting the residuals are to be collected:

$$r_i = y_i - \hat{y}_i$$

The sampling with replacement is done within the set of residuals and new observations are generated as follow:

$$y_i^{new} = \hat{y}_i + r_i^{new}, r_i^{new} \in \{r_1, r_2, ...r_N\}$$

The sample size of original data should be large enough.

For the third one, we assume that the random error are i.i.d normal:

$$y_i^{new} = \hat{y}_i + r_i^{new}, r_i^{new} \sim N(0, \sigma^2)$$

However, it is worth noting that the variance of the error terms might contain substantial heterogeneity. For example, suppose there are two states in a hidden Markov model, it is highly possible that the distributions in two state have different variance structures. Therefore, adjustments have to be applied for the method 2 and 3 (Bandeen-Roche et al. 1997, Wang et al. 2005):

1 Assign each observation to a group based on posterior probability.

2 Within each group, do the resampling of residuals.

3 Repeat 1 and 2.

By bootstrapping, new data sets can be generated and parameter estimates will be evaluated. After that, there are two ways to generate 95% confidence interval: either by $1.96 * \hat{\sigma_\beta}$ or by empirical 2.5% and 97.5% quantiles. The results should be close for sufficiently large data sets.

3.3.4 Determine the Number of Components

An important but difficult problem in hidden Markov model is to choose proper number of components. Many criteria and procedures have been proposed. In this section, we will three general approaches to handle this problem: likelihood ratio test, criteria-based model selection and cross-validation.

The likelihood ratio test (Neyman and Pearson, 1933) has been well known as an efficient way of model selection. Under certain regularity conditions, the log likelihood ratio under null hypothesis can be tested through χ^2 test. However, it has been shown that two mixture distributions with different number of components can not satisfy those regularity conditions (Wolfe, 1971). Wolfe claimed that a modified version of likelihood ratio test could be possibly applied:

$$H_0 : n = c_0$$

$$H_1 : n = c_1, (c_1 > c_0)$$

$$-\frac{2}{N}(N - 1 - d - \frac{c_1}{2})(logL(c_0) - logL(c_1)) \sim \chi^2(\frac{2d}{c_1 - c_0})$$

N is the sample size, n is the number of components, and $d = c_1 - c_0$. The assumption that $c_1 > c_0$ is based on the statistics version of "Occam's razor": We always prefer a simple model that might work. Unless there is strong evidence to support that a more complicated model is significantly better, we will stick to the simple model. Wolfe's approach is quite easy to implement in those years when computing power were quite weak. However, this approach only provides a rough approximation.

McLachlan (1987) proposed that bootstrap can be applied to obtain the approximate distribution of the log likelihood ratio test statistic under null hypothesis. The basic idea is to generate random samples by a mixture distribution with c_0 components, calculate the log likelihood ratio test statistics, and then establish the empirical distribution based on the observed test statistics. Finally, this empirical distribution can be used to calculate the p-value of original data.

The criteria-based model selection method has also gained popularity to asses the number of components in mixture models. The likelihood function, interpreted as a measure of goodness of model fitting, could not be used as a criterion to select the number of components in a mixture model due to its tendency to choose more complicated models (Biernacki et al. 2000). A number of criteria have been discussed to handle this issue. Usually these criteria add a penalty term along with the likelihood to represent the tradeoff between model complexity and utility, for example, the AIC criteria (Akaike 1974). Hurvich and Tsai (1989) suggested to use AICc (corrected AIC) instead of AIC, since AIC tends to overfit the data. The AICc

is defined as:

$$AICc = -2 * logL + 2k * (1 + \frac{(k+1)}{N-k-1})$$

k is the number of parameters.

Another popular criterion is BIC (Schwarz 1978):

$$BIC = -2 * logL + k * log(N)$$

BIC generally adds more penalty on the model complexity compared to AIC and it has been shown that BIC is equivalent to MDL (Minimum Description Length) criterion (Rissanen 1978).

Another useful criterion is Minimum Message Length (MML). MML (Wallace and Boulton 1968) was derived from the perspective of information theory. The process of modeling is considered as encoding the data and the model parameters can be considered as the extra cost of the encoding.

Therefore, the length of an encoded message can be described as:

$$Length(\theta, Y) = Length(\theta) + Length(Y|\theta)$$

If $Length(\theta)$ is short, then the model is simple but correspondingly $Length(Y|\theta)$ will be long.

An intuitive example: It has been shown that compared with the versions recorded in the other five United Nations working languages, Chinese version of a document is always the shortest (Zhao and Richard Jr 2007). If someone wants to translate the English documents

into Chinese to make it shorter, extra cost needs to be considered: a Chinese-English dictionary.

This cross-validation approach was proposed by Celeux and Durand (2008). It was based on half-sampling. Celeux showed that if we pick the odd numbers of observations or the even numbers of observations in a data set generated by hidden Markov model, the result is still a hidden Markov chain. Therefore, we may simply use the "odd subset" of the original sample to fit the model, and calculate the likelihood of the "even subset" of the original sample. The likelihood can be used as a criterion to proceed model selection.

3.3.5 Goodness of Fit

To assess the goodness of fit of a given hidden Markov model is an important topic and has drawn lots of attentions from researchers. As in regular regression models, the residuals can be used. However, due to the inherent heterogeneity of the hidden Markov model, the residuals must be adjusted by classes, as we have seen in the previous section. Wang et al. (2005) showed that the class-adjusted residuals are asymptotically equivalent to the distributions of residuals from the latent classes. Zucchini and MacDonald (2009) proposed a different approach, which is by the pseudo residual. The pseudo residual is defined as the probability of seeing a less extreme response than observed given all observations except that at time t:

$$u_t = P(Y_t \leq y_t | y_i, \forall i \neq t)$$

For well-fitted models, the pseudo residuals should be approximately U[0,1] distributed.

MacKay Altman (2004) provides an intuitive graphical approach similar to Q-Q plot. By plotting the estimated distribution against the empirical distribution, the lack of fit can be detected with high probability for large sample size. The estimated distribution is given by:

$$F(y|\theta) = \sum_{i=1}^{K} \pi_i F_i(y|\theta_i)$$

If the model is correctly specified, the plot of empirical against estimated distributions should be cloase to a staright line.

3.3.6 Prediction

Suppose the historical data have been obtained, how can we predict the travel time in the future? The basic idea follows Markov property.

Based on model specification,

$$f(y_t|s_t) = \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } s_t = 1 \\ N(\mu_2, \sigma_2^2), & \text{if } s_t = 2 \end{cases}$$

Therefore, we have:

$$E(y_t|s_t) = \begin{cases} \mu_1, & \text{if } s_t = 1 \\ \mu_2, & \text{if } s_t = 2 \end{cases}$$

If the model contains regression in the mean parameter,

$$f(y_t|s_t, x_t) = \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } s_t = 1 \\ N(\theta_0 + \theta_1 * x_t, \sigma_2^2), & \text{if } s_t = 2 \end{cases}$$

we have:

$$E(y_t|s_t, x_t) = \begin{cases} \mu_1, & \text{if } s_t = 1 \\ \theta_0 + \theta_1 * x_t, & \text{if } s_t = 2 \end{cases}$$

Given the states, the expected value of y_t can be used as the predicted value of travel time. However, the state is unobservable so our prediction is actually $E(y_t|y_1, ...y_{t-1}, x_1, ..., x_{t-1})$. The key of this problem is to predict s_t using historical data.

Assume the initial distribution for the Markov chain s_t is:

$$A = \begin{pmatrix} P(s_0 = 1) \\ P(s_0 = 2) \end{pmatrix} = \begin{pmatrix} p_0 \\ 1 - p_0 \end{pmatrix}$$

The transition matrix is:

$$T = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}$$

Then the distribution of s_1 is:

$$A * T$$

The distribution of s_2 is:

$$A * T^2$$

and so on.

The marginal distribution of s_t at any time t can be easily estimated through Markov property. The transition matrix is estimated by the data. The initial distribution could be either set manually or estimated from the last observed travel time in previous time period.

If the transition matrix is modeled through regression:

$$\log(\frac{P_{12}}{P_{11}}) = \beta_{0,1} + \beta_{1,1}x$$

$$\log(\frac{P_{22}}{P_{21}}) = \beta_{0,2} + \beta_{1,2}x$$

It is easy to see that:

$$P_{11} = \frac{exp(\beta_{0,1} + \beta_{1,1}x)}{exp(\beta_{0,1} + \beta_{1,1}x) + 1}$$

$$P_{12} = \frac{1}{exp(\beta_{0,1} + \beta_{1,1}x) + 1}$$

$$P_{21} = \frac{exp(\beta_{0,2} + \beta_{1,2}x)}{exp(\beta_{0,2} + \beta_{1,2}x) + 1}$$

$$P_{22} = \frac{1}{exp(\beta_{0,2} + \beta_{1,2}x) + 1}$$

Consider a simple example: Suppose that during the time interval [7:00-7:59], the traffic volume is 8. The transition matrix is:

$$\log(\frac{P_{12}}{P_{11}}) = -6 + 0.1 * x$$

$$\log(\frac{P_{22}}{P_{21}}) = 0.6 + 0.15 * x$$

$$T(8) = \begin{pmatrix} 0.9946 & 0.0054 \\ 0.8581 & 0.1419 \end{pmatrix}$$

Since the distribution of first vehicle is unknown, we might use the non-informative prior:

$$A = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

If we want to predict the state of the first vehicle in the next time interval, i.e. s_8, then it can be showed that:

$$A * T(8)^7 = \begin{pmatrix} 0.994 \\ 0.006 \end{pmatrix}$$

That is, y_8 has 99.4% probability to be in the free-flow state and μ_1 can be used as predicted value.

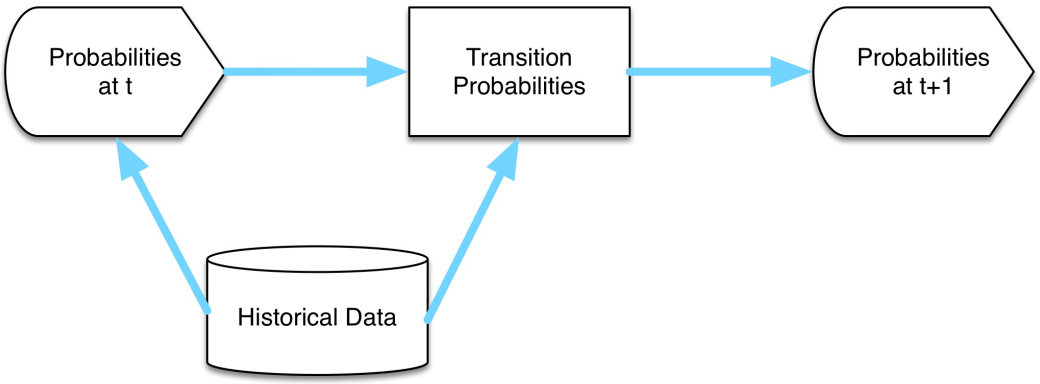

Figure 3.7: Hidden Markov Model: Estimation

Figure 3.7 is an overview of the prediction procedures in hidden Markov model.

3.4 Simulation Study

3.4.1 No Covariate

First we consider a simple case when the covariate is not present in the model. We simulate 1000 data sets, each with 5000 observations, according to the transition matrix: (values are based on estimates from real data)

$$\begin{pmatrix} 0.992 & 0.008 \\ 0.068 & 0.932 \end{pmatrix}$$

Each data set will be fitted by both traditional and hidden Markov models. Table 3.1 implies that the interval estimates of HMM are slightly narrower.

Table 3.1: HMM Vs. Traditional: No Covariate 1

Name	True	Traditional	95% C.I.	HMM	95% C.I.
μ_1	580	579.9	(578.4, 581.2)	579.9	(578.4, 581.1)
μ_2	1035	1037.1	(989.8, 1067.1)	1036.6	(1001.2,1062.5)
σ_1	41	40.9	(40.0, 42.1)	40.9	(40.1, 42.1)
σ_2	371	369.8	(348.4, 389.0)	370.1	(350.3, 388.9)

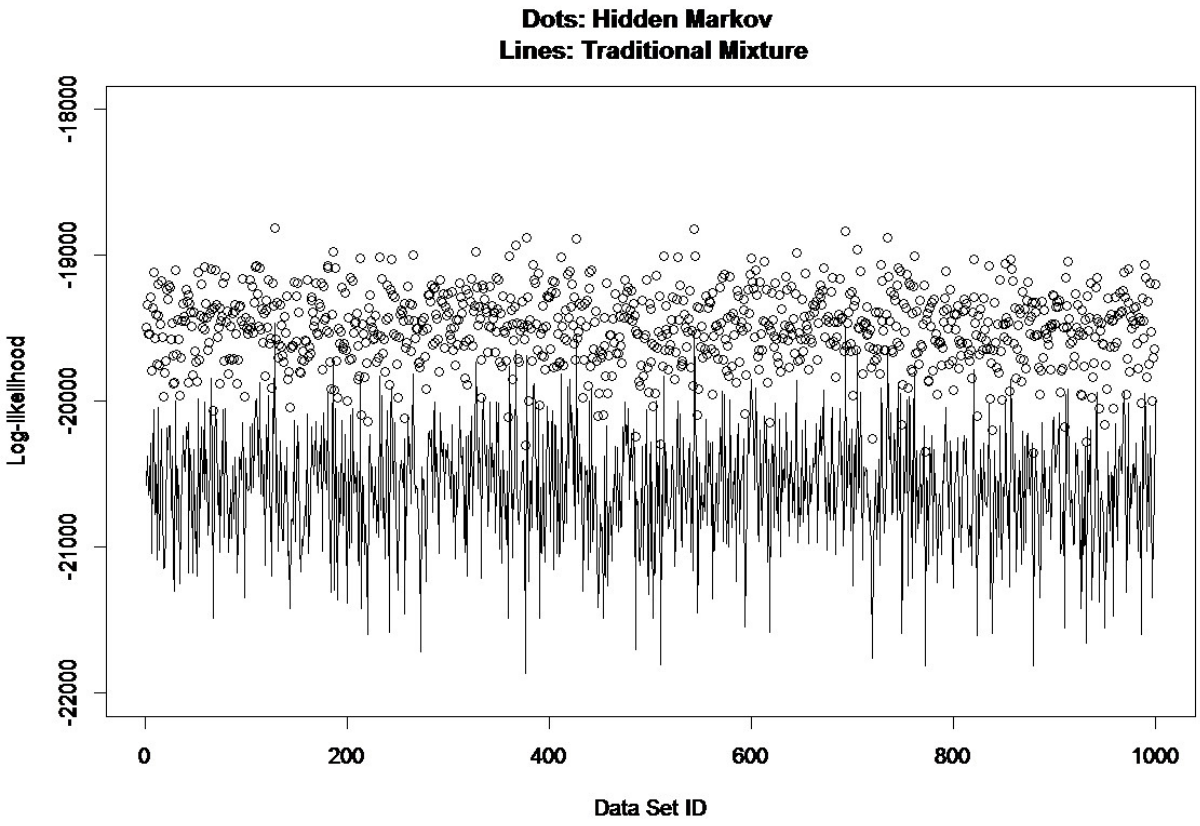

Figure 3.8: HMM Vs. Traditional 1

Figure 3.8 indicates that the log-likelihood of HMM models are much larger than that of traditional models. The mean difference in log-likelihood is around 951, which is a huge difference.

We also tried another set of parameters and Table 3.2 also implies that HMM can generate slightly narrower confidence intervals.

Table 3.2: HMM Vs. Traditional: No Covariate 2

Name	True	Traditional	95% C.I.	HMM	95% C.I.
μ_1	580	579.9	(578.7,581.0)	579.8	(578.8,581.2)
μ_2	750	751.6	(708.1,793.4)	751.2	(710.1,788.0)
σ_1	41	41.0	(40.1, 42.1)	41.0	(40.0 41.9)
σ_2	371	369.2	(341.3,396.3)	369.2	(340.6,394.0)

3.4.2 With Covariate

When the covariate is considered in the model, we also showed that HMM is superior to traditional mixture model. We simulate 500 data sets, each with 5000 observations, according to the parameters setting: (values are based on estimates from real data):

$$log(y) \sim \begin{cases} N(log(500), \sigma_1 = 0.07), & \text{if } s_t = 1 \\ N(log(1000), \sigma_2 = 0.31), & \text{if } s_t = 2 \end{cases}$$

$$\log(\frac{P_{12}}{P_{11}}) = -6 + 0.1 * x$$

$$\log(\frac{P_{22}}{P_{21}}) = 0.6 + 0.15 * x$$

Due to computing issues, we use log transform of the original data and the log likelihood values will be changed accordingly.

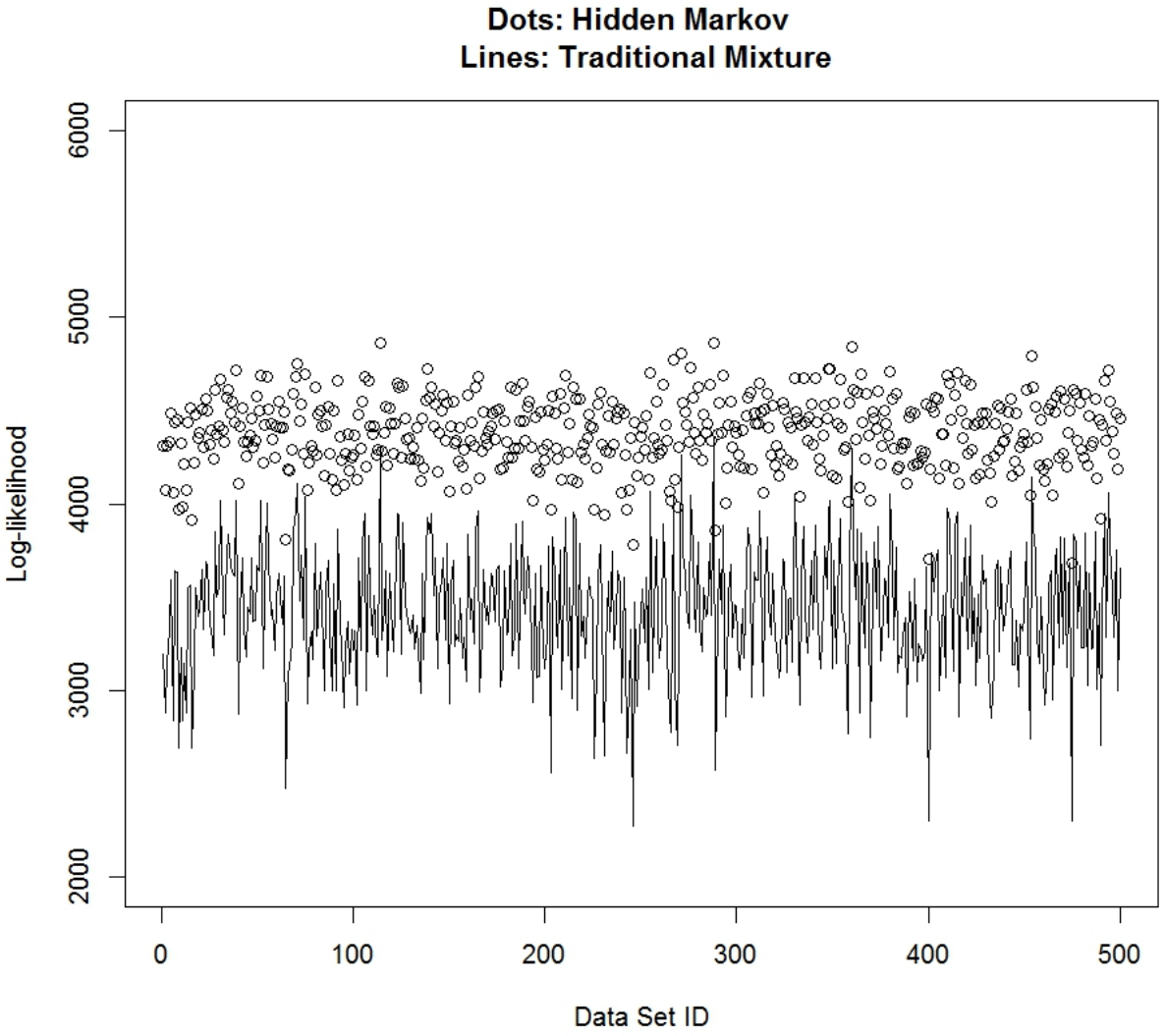

Figure 3.9: HMM Vs. Traditional 2

Figure 3.9 indicates that the log-likelihood of HMM models are much larger than that of traditional models. The mean difference in log-likelihood is around 997, which is a huge difference.

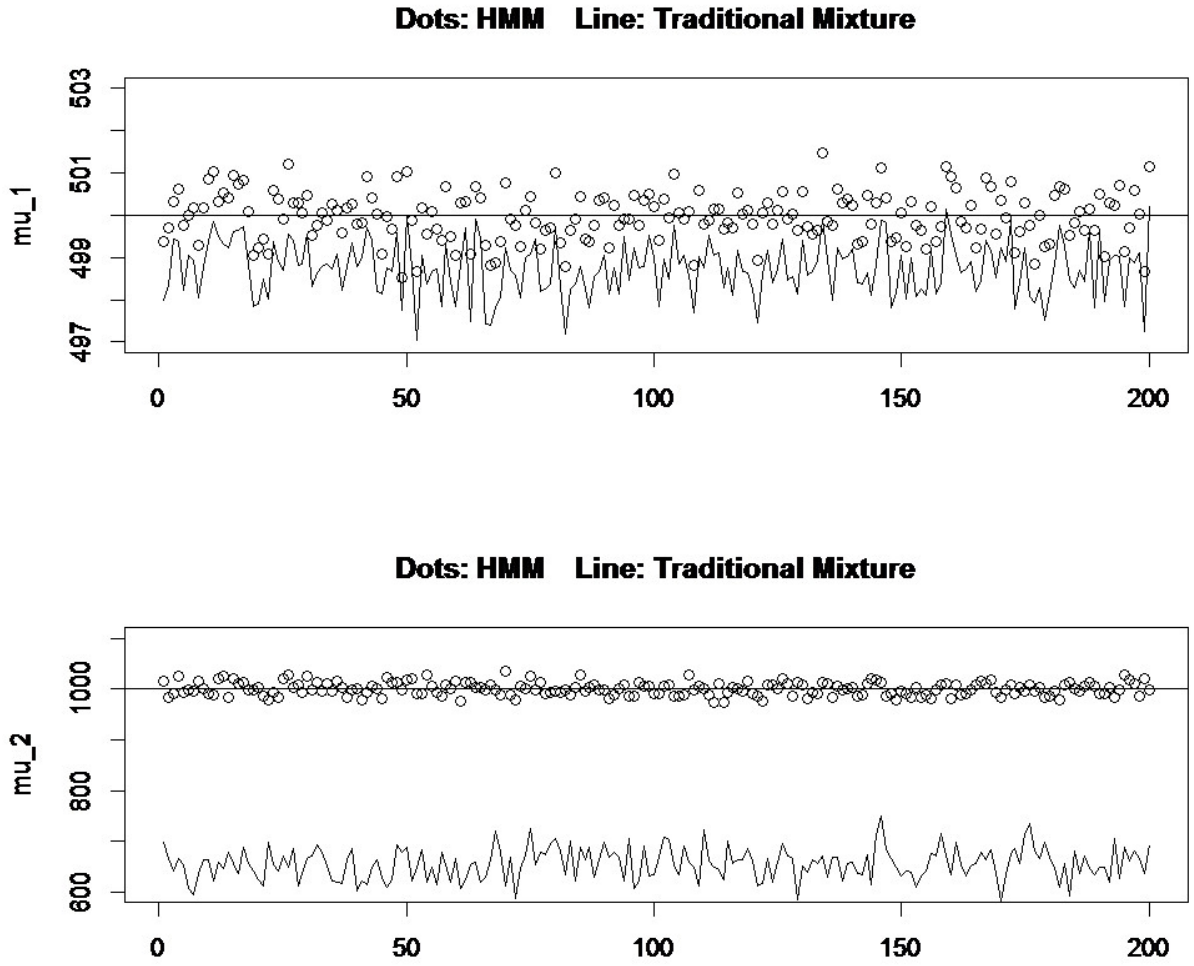

Figure 3.10: HMM Vs. Traditional 3

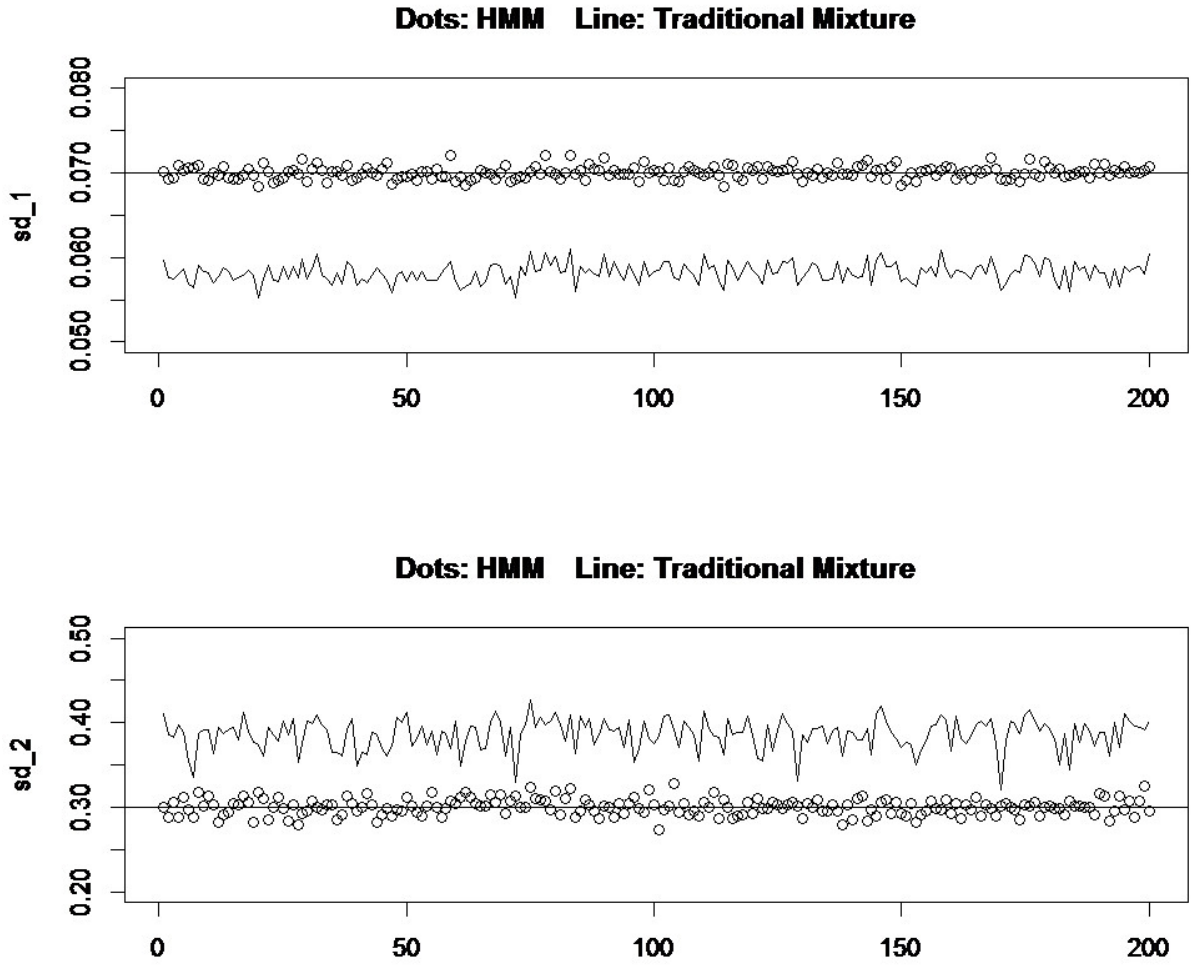

Figure 3.11: HMM Vs. Traditional 4

Figure 3.10 and 3.11 clearly demonstrate the advantage of HMM. Both the mean estimates and the variance estimates from HMM are superior.

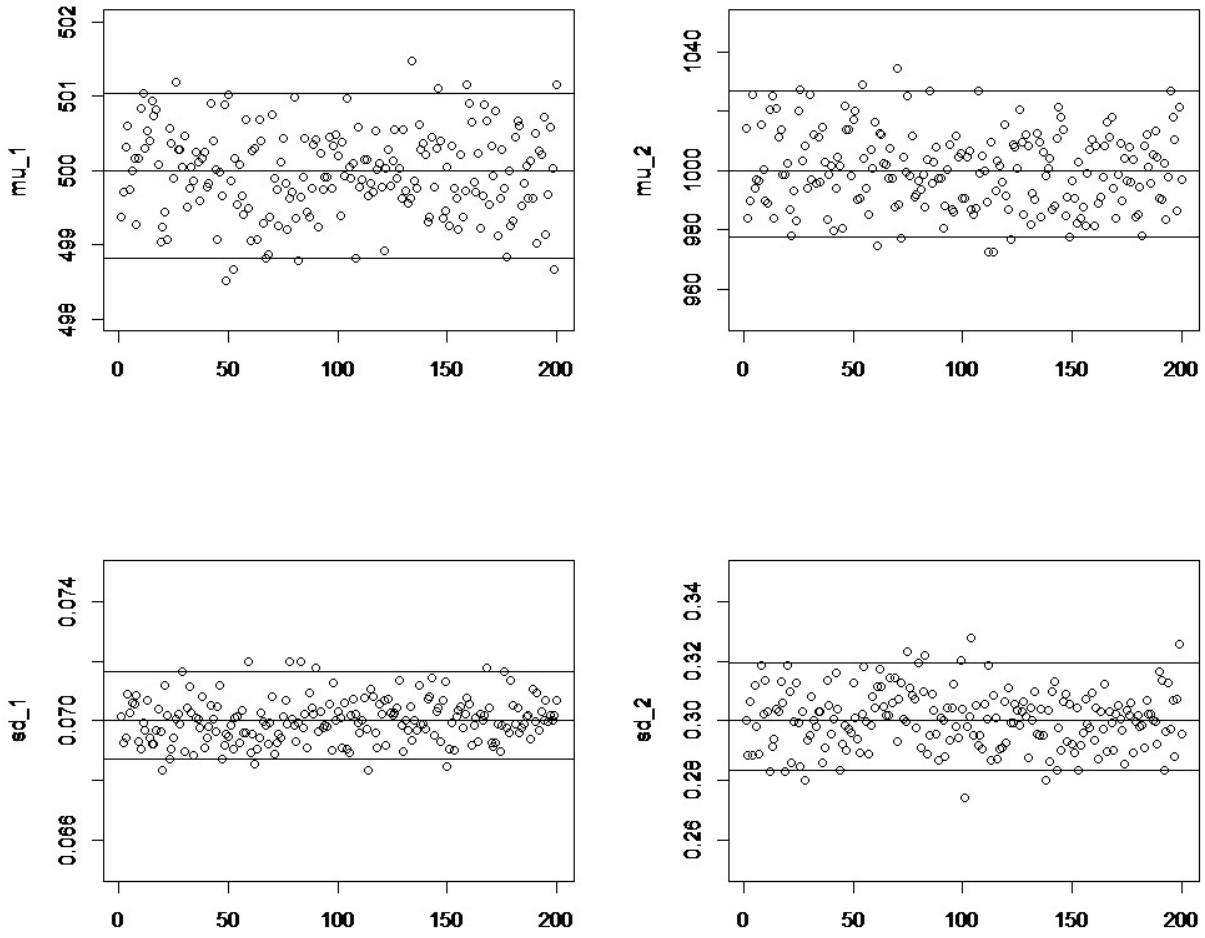

Figure 3.12: 95% C.I. of HMM

Figure 3.12 illustrates the 95% confidence intervals of several parameters estimates of hidden markov model. The estimates from different samples are relatively symmetric and centered at the true values.

Table 3.3: Parameter Estimation of HMM

Name	True	95% C.I.
μ_1	500	(499,501)
μ_2	1000	(977.7,1026.7)
σ_1	0.07	(0.070, 0.071)
σ_2	0.3	(0.28,0.32)
$\beta_{0,1}$	-6	(-7.04,-5.07)
$\beta_{1,1}$	0.1	(0.03,0.16)
$\beta_{0,2}$	0.6	(-0.56,1.52)
$\beta_{1,2}$	0.15	(0.09, 0.23)

Table 3.3 provides the numbers in Figure 3.12.

3.5 Results for Real Data

The data collected in I-35 Highway represents the actual traffic flow by a scale. Only the vehicle equipped with electronic devices were counted.

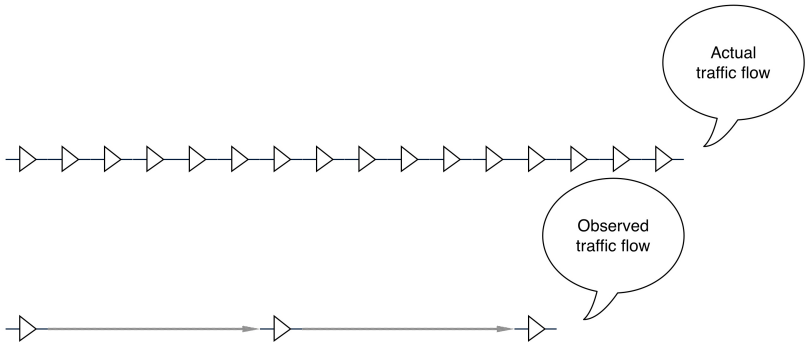

Figure 3.13: Illustration of Low Sampling Rate

Figure 3.13 illustrates that sampling rate was realtively low and the observations should be considered as a small proportion of the actual traffic flow.

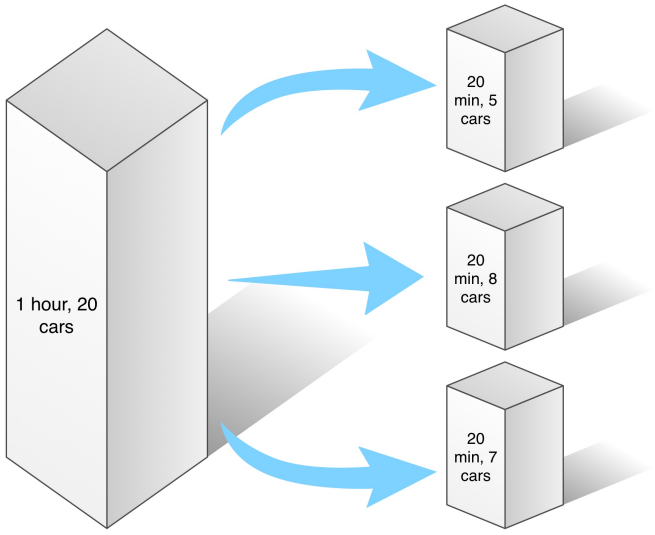

Figure 3.14: Illustration of Potential Improvement

In order to predict the travel time with higher precision, the sampling rate could be increased (Figure 3.14) but the basic modeling steps are still the same.

The number of hidden states in real data can be determined through likelihood ratio test. We will consider that if two states are sufficient to depict the hidden structure the data. Otherwise we will move forward to three states:

$$H_0 : n = 2$$

$$H_1 : n = 3$$

Since the log likelihood ratio does not follow χ^2 distribution, we have to consider bootstrap sampling strategy. Figure 3.15 is the histogram of the log likelihood ratio from 500 samples.

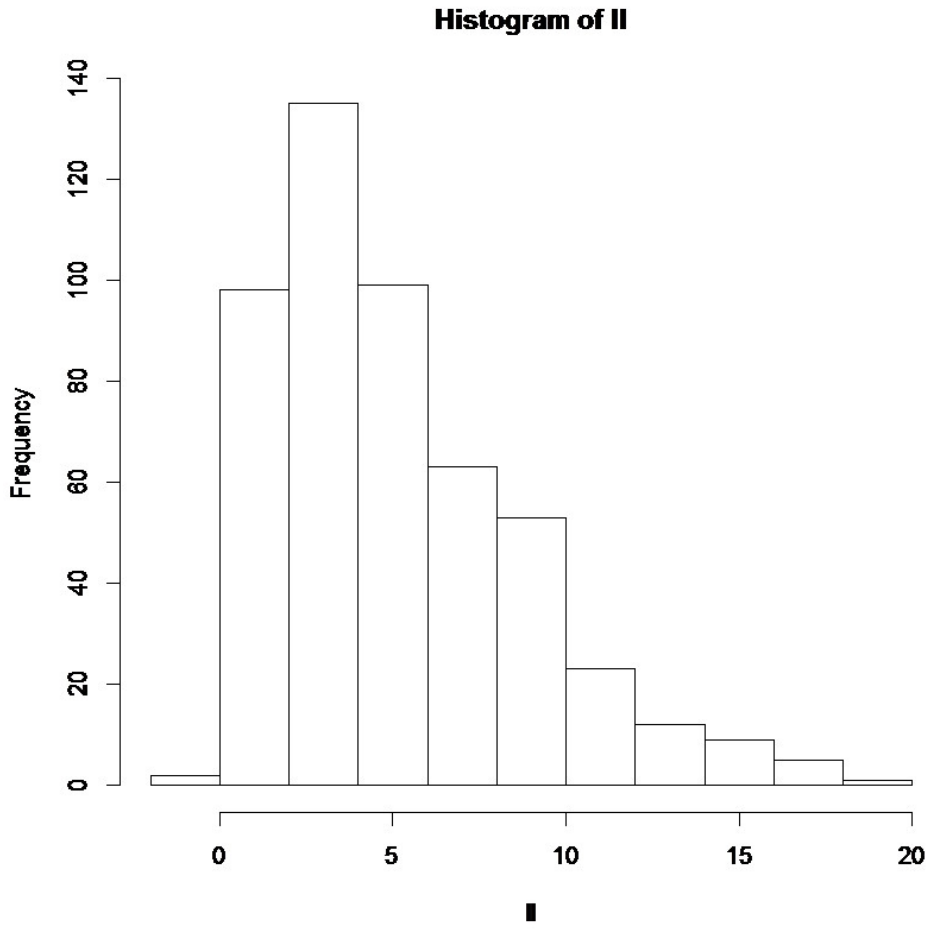

Figure 3.15: Histogram of the Log Likelihood Ratio

The observed log likelihood ratio has the value around 577.2, whose p-value is significantly smaller than 0.05. Therefore, er reject the null hypothesis.

Although the empirical distribution may not be χ^2, it is interesting to assess the deviance between them. I use Kolmogorov-Smirnov test to test the null hypothesis that the empirical distribution follows χ^2 with a set of different degrees of freedom. The results is shown in Table 3.4

Table 3.4: Kolmogorov-Smirnov Test Result

Degree of Freedom	P-value
4	1.598e-10
5	0.03336
6	1.788e-10
7	$<$ 2.2e-16

It is easy to see that none of these χ^2 distributions is good enough to describe the empirical distribution. The closet one is $\chi^2(5)$. Figure 3.16 compares the two and the empirical distribution has heavier mass on the left (heavier tail). This implies that, suppose the null hypothesis $k = 2$ is true, the χ^2 test will be more likely to reject it than the bootstrap one, which yields Type-I error. However, if the null hypothesis is false, the two tests have similar power to reject it.

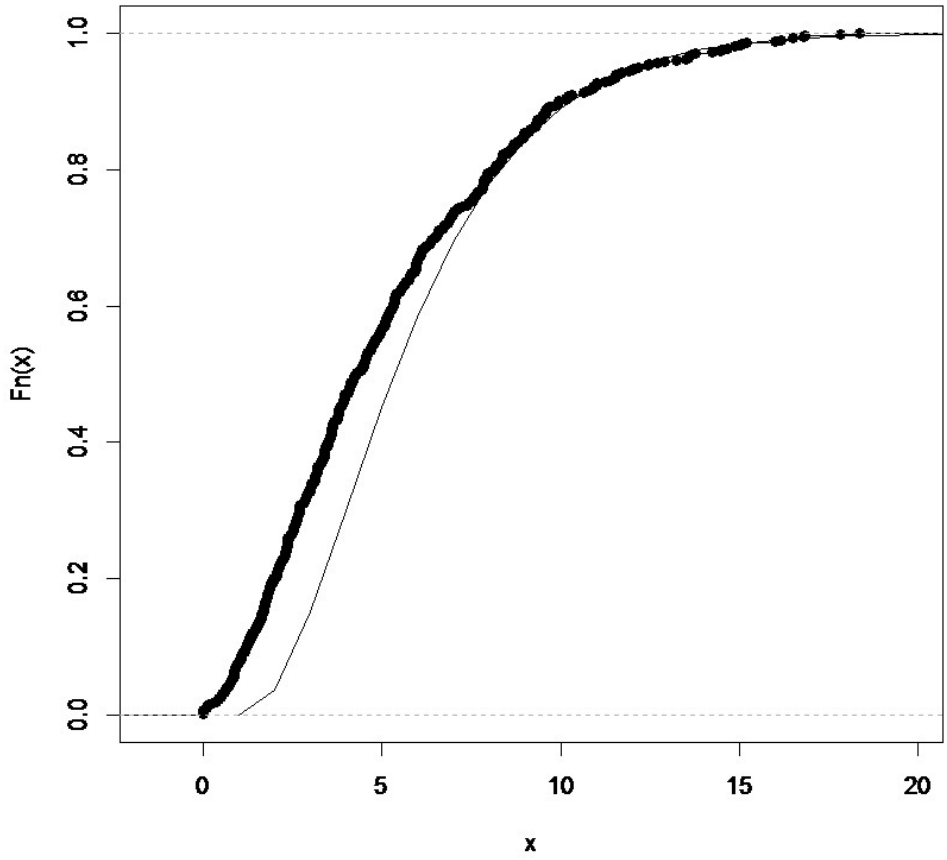

Figure 3.16: χ^2 and Empirical Distributions

The BIC for two-component and three-component models are -8572.2 and -9090.8 respectively. This also implies that three-component model is superior. For the Message Length criterion, the values are -4280.6 and -4548.8. We also tried AICc and similar results were obtained. Based on the half-sampling cross validation, the log likelihoods are 2066.8 and 2178.6. In sum, all of the results indicated that actually three-component model are better.

The three states model is:

$$\log(\frac{P_{12}}{P_{11}}) = \beta_{0,1} + \beta_{1,1}x \qquad \log(\frac{P_{13}}{P_{11}}) = \beta_{0,2} + \beta_{1,2}x$$

$$\log(\frac{P_{22}}{P_{21}}) = \beta_{0,3} + \beta_{1,3}x \qquad \log(\frac{P_{23}}{P_{21}}) = \beta_{0,4} + \beta_{1,4}x$$

$$\log(\frac{P_{32}}{P_{31}}) = \beta_{0,5} + \beta_{1,5}x \qquad \log(\frac{P_{33}}{P_{31}}) = \beta_{0,6} + \beta_{1,6}x$$

$$f(y_t|s_t) = \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } s_t = 1 \\ N(\mu_2, \sigma_2^2), & \text{if } s_t = 2 \\ N(\mu_3, \sigma_3^2), & \text{if } s_t = 3 \end{cases}$$

$$\mu_i = \theta_{0,i} + \theta_{1,i}$$

Table 3.5: Parameter Estimation for Real Data

$\beta_{0,1}$	$\beta_{1,1}$	$\beta_{0,2}$	$\beta_{1,2}$	$\beta_{0,3}$	$\beta_{1,3}$
-4.89	0.056	-5.37	0.383	1.09	0.063
$\beta_{0,4}$	$\beta_{1,4}$	$\beta_{0,5}$	$\beta_{1,5}$	$\beta_{0,6}$	$\beta_{1,6}$
-2.24	0.10	-1.3	0.25	0.46	0.33
$\theta_{0,1}$	$\theta_{1,1}$	$\theta_{0,2}$	$\theta_{1,2}$	$\theta_{0,3}$	$\theta_{1,3}$
6.35	0*	6.46	0.005	7.03	0*
σ_1	σ_2	σ_3			
0.066	0.092	0.27			

* Note: 0 means not significant.

In general, when the traffic volume is higher, the free-flow state will be more likely to move

to congested state and the congested state will be more likely to stay. It is worth noting that there is a medium state within free-flow and congested. The average value of P_{31} is around 10^{-14}, so it is very unlikely to move from congested to free-flow directly. Most of the time the chain will use medium state as "interim period", as shown in Figure 3.17.

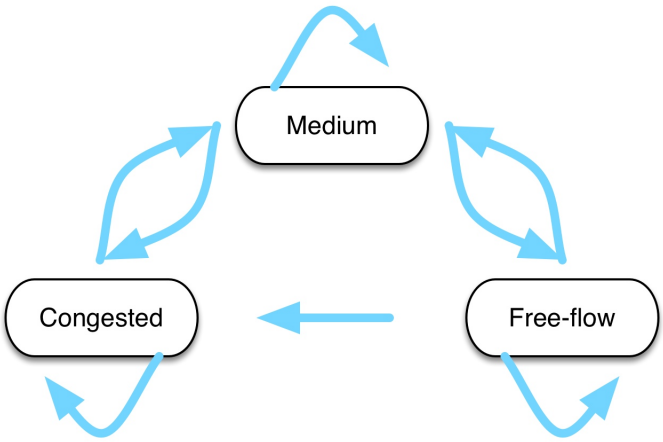

Figure 3.17: Illustration of Three States Markov Chain

The marginal distribution for the three states estimated by Viterbi algorithm is: Free-flow 84.4%, Medium 8.2%, Congested 7.4%.

The pseudo class adjusted residual plots indicate that the free-flow and medium states are really close to normal while the congested state is slightly skewed. The standardized residuals

plot implies that the residuals are generally within the range $(-3, 3)$ and do not significantly increase by time.

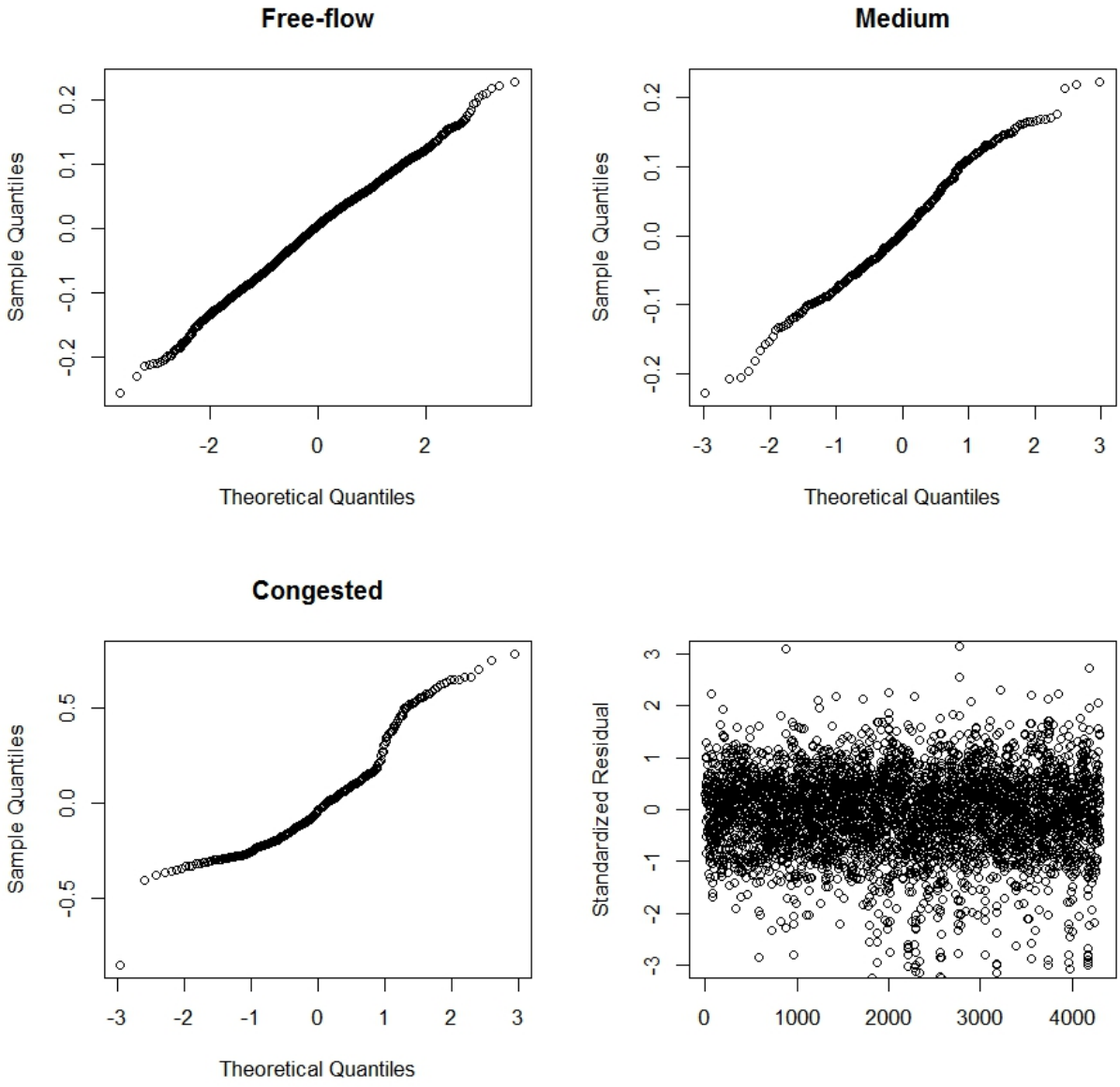

Figure 3.18: Residual Check

3.6 Summary

We apply the hidden Markov model to the travel time reliability problem in order to accommodate the dependency structure of observations and to understand how the traffic volume influences on the travel time of vehicles.

Regarding the model specification we consider two possible states of the hidden Markov chain, which are "free-flow" and "congested". The parameters and proportions of the two states are estimated. Moreover, we apply the well-known logit function in the transition matrix to include the covariate of traffic volume. The modeling result shows that the traffic volume has a positive effect on the proportion of "congested" condition as well as the mean parameters of such condition.

We have compared the model fitting of hidden Markov model with that of ordinary mixture model, and the significant improvement has been shown. To sum up, the hidden Markov model is superior to interpret the data without sacrificing model simplicity.

Chapter 4

Individual Driver Risk: Poisson Mixture and Overlap Probability

4.1 Introduction

In this chapter we will discuss about the problem of clustering safe and risky drivers. The general form of a mixture Poisson distribution density function is:

$$f(x) = \int_0^{+\infty} f_P(x|\lambda) g(\lambda) d\lambda,$$

where $f_P(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2...$ and $g(\lambda)$ is a density function whose support is constrained in nonnegative real numbers.

The n-component Poisson distribution is a special case of mixture Poisson distribution s.t.

$g(\lambda) = \sum_{i=1}^{n} \alpha_i \delta_{\lambda_i}(\lambda)$, where $\sum_{i=1}^{n} \alpha_i = 1$, $\alpha_i > 0$ and

$$\delta_{\lambda_i}(\lambda) = \begin{cases} 1 & \text{if } \lambda = \lambda_i \\ 0 & \text{otherwise} \end{cases}$$

The n-component Poisson distribution density, denoted by $f_{P(n)}$, has the following form:

$$f_{P(n)}(x|\alpha_1, ..., \alpha_n, \lambda_1, ..., \lambda_n) = \sum_{i=1}^{n} \alpha_i f_P(x|\lambda_i)$$

It is worth noting that under the constraints $\sum_{i=1}^{n} \alpha_i = 1$, the n-component Poisson model has only $2n - 1$ free parameters.

The standard Poisson distribution has an important property: the mean is equal to the variance, which is usually violated by the overdispersion in transportation data. However, the n-component Poisson distribution is not constrained by this:

$$E(X) = \sum_{i=1}^{n} \alpha_i \lambda_i$$

$$Var(X) = E(X^2) - (E(X))^2 = \sum_{i=1}^{n} \alpha_i(\lambda_i^2 + \lambda_i) - (\sum_{i=1}^{n} \alpha_i \lambda_i)^2$$

The variance is greater than the mean for n-component Poisson distribution, with the only exception that the mixture degenerates to a standard Poisson. Therefore, the n-component Poisson model is able to accommodate overdispersion data. In general, if $X \sim f_{P(n)}(x|\alpha_1, ..., \alpha_n, \lambda_1, ..., \lambda_n)$, then $Var(X) \geq E(X)$. The equality holds if and only if $\lambda_i = \lambda_1$ for $\forall i$.

The proof is very simple. Denote $g(x) = x^2$, which is a convex function. According to Jensen's inequality, $\sum_{i=1}^{n} \alpha_i \lambda_i^2 - (\sum_{i=1}^{n} \alpha_i \lambda_i)^2 \geq 0$.

Without loss of generality, we always assume $\lambda_j > \lambda_i$ if $j > i$. Before yhe discussion of mixture Poisson, we will introduce the overlap probability as a measure for two individual Poisson distributions.

The simplest n-component Poisson distribution is a two-component Poisson model, whose probability density function is the weighted average of two Poisson density functions. For simplicity, we will use a slightly different parameterization: $f_{P(2)}(x|\alpha, \lambda_1, \lambda_2) = \alpha f_P(x|\lambda_1) + (1-\alpha) f_P(x|\lambda_2)$.

The mixture Poisson distribution is usually used as a model-based clustering technique for count data. Suppose a population contains two groups of observations, both follow Poisson distribution but with different parameters λ_1 and λ_2.

The plot below shows an example of two-component Poisson distribution, which is denoted by the curve with dots. The density function is: $f_{P(2)}(x|\alpha = 0.5, \lambda_1 = 2, \lambda_2 = 10) = 0.5 * f_P(x|2) + 0.5 * f_P(x|10)$. Two standard Poisson density curves are also overlayed in the Figure 2 but neither can sufficiently represent the shape of the mixture. Actually, the mixture implies a bimodal shape, which is different from any standard Poisson distributions.

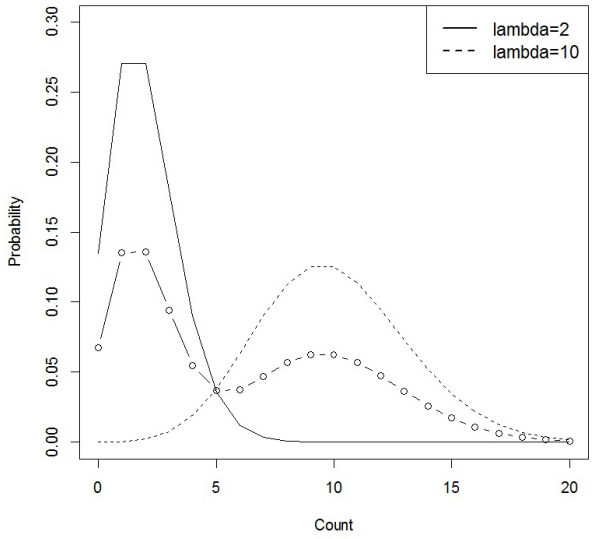

Figure 4.1: Illustration of Two-Component Poisson

4.2 Data Aggregation and the Impact

It is well known that traffic accidents are rare event for individual driver (Ng et al. 2002, Formisano et al. 2005). Therefore, data aggregation is a common technique in the analysis of driver risk. For example, in a single year, most of drivers might commit zero accidents, which makes the safe and risky groups not well separated. If the consecutive ten years crash data are collected, the total number of crashes during the entire time period for each driver can be summarized, and consequently the safe and risky groups could be separated better.

However, there are also some drawbacks in data aggregation. Schuh et al. (2013) discussed about the impact of data aggregation from the perspective of quality control, and claimed

that the data aggregation could lead to potential information loss because of a delay in detecting an increased risk. This drawback is not so essential in driver risk analysis since the data collected for each year satisfy independent condition.

Although at the first glance the independent assumption mighr be too strong, for the CEI data set we calculated the correlation among number of accidents for each years and the results were uniformly smaller than 0.1. Thus, the independent assumption should be reasonable.

There is another potential issue of the data aggregation. It is well known that the summation of independent Poisson distributed variables are still Poisson. On the other hand, if the sum of two independent non-negative random variables X and Y has a Poisson distribution, then both X and Y themselves must have the Poisson distribution (Raikov 1937). But the summation of independent mixture Poisson distributed variables may no longer be mixture Poisson. We will show this issue by moment generating function.

The moment generating function (MGF) is defined as:

$$M_X(t) = E[e^{tX}]$$

For Poisson distributed random variable with parameter λ, the MGF is:

$$e^{\lambda(e^t-1)}$$

For two-component mixture Poisson distributed random variable with density $\alpha f_P(x|\lambda_1) +$

$(1-\alpha)f_P(x|\lambda_2)$, the MGF is:

$$\alpha e^{\lambda_1(e^t-1)} + (1-\alpha)e^{\lambda_2(e^t-1)}$$

One of the important property of MGF is: If x and y are independent random variables, then the MGF of $x+y$ will be the product of MGF's for x and y. Suppose x and y are i.i.d Poisson distributed with parameters λ, then it can be shown that:

$$e^{\lambda(e^t-1)} * e^{\lambda(e^t-1)} = e^{2\lambda(e^t-1)}$$

Therefore, x+y follows a Poisson distribution with parameter 2λ. The mixture Poisson does not hold this property:

$$(\alpha e^{\lambda_1(e^t-1)} + (1-\alpha)e^{\lambda_2(e^t-1)}) * (\alpha e^{\lambda_1(e^t-1)} + (1-\alpha)e^{\lambda_2(e^t-1)})$$

$$= \alpha^2 e^{2\lambda_1(e^t-1)} + 2\alpha(1-\alpha)e^{(\lambda_1+\lambda_2)(e^t-1)} + (1-\alpha)^2 e^{2\lambda_2(e^t-1)}$$

The product no longer follows a mixture Poisson MGF form. The plot below clearly illustrates this phenomenon:

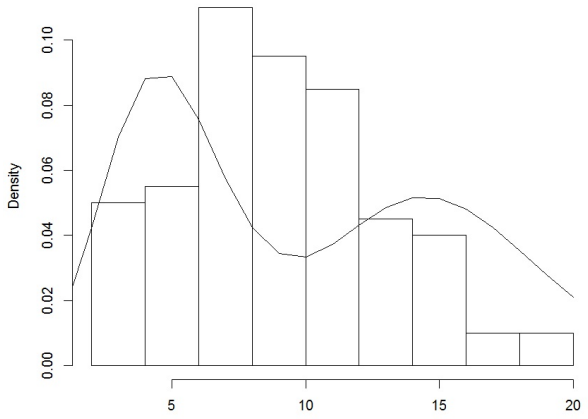

Figure 4.2: Summation of Mixture Poisson $\neq$ Mixture Poisson

The histogram is generated by the sum of five independence samples of mixture Poisson. On the other hand, the curve shows the mixture Poisson density assuming that the mean parameters can be added up. The sum of independent mixture Poisson random variables has a unimodal shape, which is not surprising if we recall the central limiting theorem.

The data aggregation is valid under the assumption that the group (safe or risky) each drivr belongs to is fixed and known, and the number of accidents in each year is independent for each driver. For example, assuming that this driver is from safe group, which can be characterized by parameter λ_1, then under the independent assumption we may claim that the number of accidents for this driver in $[0, 2t]$ will follow a $Poisson(2\lambda_1)$.

Baetschmann and Winkelmann (2013) included "exposure" as the covariate in zero-inflated model. That is, the time itself will be used in the model:

$$\alpha(t) f_P(x|\lambda_1(t)) + (1 - \alpha(t)) f_P(x|\lambda_2(t))$$

However, this approach requires that the exposure (driver's mileage) has been accurately measured, which is rarely the case in real data set. Therefore, we will still follow the ordinary data aggregation procedure. The plot below is a conceptual illustration of data aggregation.

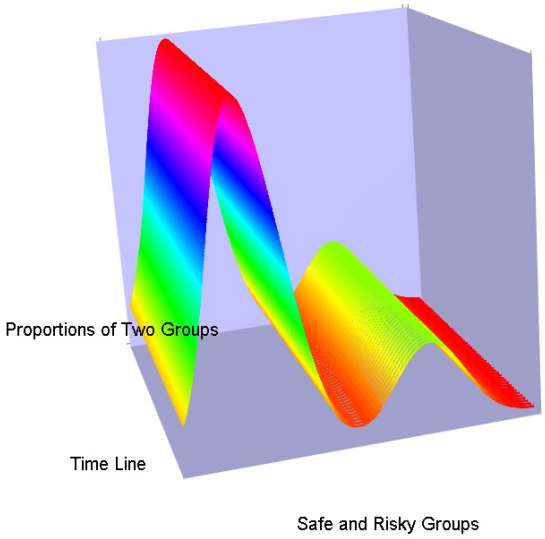

Figure 4.3: Conceptual Illustration of Data Aggregation

4.3 Model Estimation

4.3.1 EM Algorithm

A common way to estimate the mixture Poisson model is the EM algorithm. The EM algorithm has some potential drawbacks. For example, it is sensitive to the initial value specification and might be trapped in the local maxima rather than global one. However, it is still a popular choice due to the simplicity of its implementation.

E step:

$$\pi_{i1} = P(x_i \in Group_1|.) = \frac{f_P(x_i|\hat{\lambda_1})\pi_1}{f_P(x_i|\hat{\lambda_1})\pi_1 + f_P(x_i|\hat{\lambda_2})\pi_2}$$

$$\pi_{i2} = P(x_i \in Group_2|.) = 1 - \pi_{i1}$$

M step:

$$\hat{\lambda_1} = \frac{\sum_{i=1}^{n} \pi_{i1} * x_i}{\sum_{i=1}^{n} \pi_{i1}} \qquad \hat{\lambda_2} = \frac{\sum_{i=1}^{n} \pi_{i2} * x_i}{\sum_{i=1}^{n} \pi_{i2}}$$

$$\pi_1 = \hat{\alpha} = \frac{\sum_{i=1}^{n} \pi_{i1}}{n} \quad \pi_2 = 1 - \hat{\alpha} = \frac{\sum_{i=1}^{n} \pi_{i2}}{n}$$

4.3.2 Bayesian Analysis: No Covaiate

For the Bayesian analysis without covariates, it has been discussed by Dellaportas et al. (1997). Suppose the data set Y comes from k-component Poisson distribution.

$$Y = \{y_i, i = 1, 2, ..., n\} \sim f_{P(k)}(y|\lambda_i, ..., \lambda_k) = \sum_{j=1}^{k} p_j f_P(y_i|\lambda_j)$$

where $f_P(y|\lambda_j) = \frac{e^{-\lambda_j}\lambda_j^y}{y!}, y = 0, 1, 2..., j = 1, 2, ..., k.$ $p_j > 0$ and $\sum_{j=1}^{k} p_j = 1$.

For the convenience of Bayesian inference, the indicator variables z_{ij} could be used in the model.

$$z_{ij} = \begin{cases} 1 & \text{if } y_i \sim f_P(\lambda_j) \\ 0 & \text{otherwise} \end{cases}$$

$$f(z_{.j} = 1|\mathbf{p}) = p_j$$

By introducing z_{ij}, the likelihood for y_i becomes: $\prod_{j=1}^{k}(f_P(y_i|\lambda_j))^{z_{ij}}$

According to Bayesian theorem, the joint posterior distribution of the parameters is:

$$f(\mathbf{p}, \lambda, \mathbf{z}|\mathbf{y}) \propto f(\mathbf{y}|\mathbf{p}, \lambda, \mathbf{z})\pi(\mathbf{p}, \lambda, \mathbf{z})$$

$$= f(\mathbf{y}|\lambda, \mathbf{z})\pi(\mathbf{z}|\mathbf{p})\pi(\mathbf{p})\pi(\lambda)$$

The following prior distribution were used:

$$\pi(\mathbf{z_i}|\mathbf{p}) \sim Multinomial(1, \mathbf{p})$$

$$\pi(\lambda_j) \sim Gamma(a_j, b_j)$$

$$\pi(\mathbf{p}) \sim Dirichlet(\mathbf{d})$$

We can use Gibbs sampler to obtain the joint posterior distributions:

$$f(\mathbf{p}|.) \propto \prod_{j=1}^{k} p_j^{d_j-1} \prod_{i=1}^{n}\prod_{j=1}^{k} p_j^{z_{ij}} = \prod_{j=1}^{k} p_j^{d_j+\sum_{i=1}^{n} z_{ij}-1}$$

$$\Rightarrow f(\mathbf{p}|.) = Dirichlet(d_1 + \sum_{i=1}^{n} z_{i1}, ..., d_k + \sum_{i=1}^{n} z_{ik})$$

$$f(\lambda|.) \propto \prod_{j=1}^{k} e^{-b_j\lambda_j}\lambda_j^{a_j-1} \prod_{i=1}^{n}\prod_{j=1}^{k} f_P(y_i|\lambda_j)^{z_{ij}} = \prod_{j=1}^{k} e^{-\lambda_j(b_j+\sum_{i=1}^{n} z_{ij})}\lambda_j^{a_j+\sum_{i=1}^{n} y_i z_{ij}-1}$$

$$\Rightarrow f(\lambda_j|.) = Gamma(a_j + \sum_{i=1}^{n} y_i z_{ij}, b_j + \sum_{i=1}^{n} z_{ij}) * I(\lambda_{j-1} < \lambda_j < \lambda_{j+1})$$

$$f(\mathbf{z}|.) \propto \prod_{i=1}^{n}\prod_{j=1}^{k} f_P(y_i|\lambda_j)^{z_{ij}} p_j^{z_{ij}}$$

$$\Rightarrow f(\mathbf{z_{i.}}|.) = Multinomial(1, w_1, w_2, ..., w_k), \text{where } w_j = \frac{f_P(y_i|\lambda_j)p_j}{\sum_{j=1}^{k} f_P(y_i|\lambda_j)p_j}.$$

4.3.3 Bayesian Analysis: Regression on Rates

The simple mixture Poisson model only concerns about the response, while sometimes the data set also contains the useful information such as age, gender and exposure that might

help model fitting. If we assume that the crash rate depends on some covariates to be estimated, then the problem will become something similar to "generalized linear model" (GLM).

However, the n-component Poisson distribution is not an exponential family distribution given that p_j's and λ_j's are unknown (Akaho 2008). Therefore, the standard software packages may not be used. Wang et al. (1996), Yang and Lai (2005) and Wedel et al. (1993) estimated the parameters in the model through EM algorithm. It is easy to show that the probit link can be used and a Bayesian model speicification is as follow:

$$y_i \sim \sum_{j=1}^{k} p_j f_P(x_i|\lambda_{ij})$$

$$log(\lambda_{ij}) = X_i\beta_j$$

y_i: Observed number of crashes for subject i.

X_i: Covariate matrix for subject i (include intercept).

β_j: The parameters for subjects in Group j.

z_{ij}, p_j: As the definition in 4.3.2.

Each subgroup has its own set of parameters, while in traditional Poisson regression $\beta_j = \beta, \forall j$. A logarithm function is used.

Assume $\pi(\beta) \sim N(\mu_{\mathbf{0}}, \Sigma_0)$

$$f(\mathbf{p}|.) \propto \prod_{j=1}^{k} p_j^{d_j-1} \prod_{i=1}^{n}\prod_{j=1}^{k} p_j^{z_{ij}} = \prod_{j=1}^{k} p_j^{d_j+\sum_{i=1}^{n} z_{ij}-1}$$

$$\Rightarrow f(\mathbf{p}|.) = Dirichlet(d_1 + \sum_{i=1}^{n} z_{i1}, ..., d_k + \sum_{i=1}^{n} z_{ik})$$

$$f(\beta|.) \propto N(\mu_{\mathbf{0}}, \Sigma_0) \prod_{i=1}^{n}\prod_{j=1}^{k} (\frac{e^{-e^{X_i\beta_j}}}{y_i!} e^{X_i\beta_j y_i})^{z_{ij}}$$

$$f(\mathbf{z}|.) \propto \prod_{i=1}^{n}\prod_{j=1}^{k} f_P(y_i|e^{X_i\beta_j})^{z_{ij}} p_j^{z_{ij}}$$

$$\Rightarrow f(\mathbf{z_{i.}}|.) = Multinomial(1, w_1, w_2, ..., w_k), \text{where } w_j = \frac{f_P(x_i|e^{X_i\beta_j})p_j}{\sum_{j=1}^{k} f_P(x_i|e^{X_i\beta_j})p_j}.$$

The full conditional distribution for β does not have closed form, hence Metropolis-Hasting algorithm is required.

4.3.4 Bayesian Analysis: Regression on Proportion

We adopted an alternative way to include the covariates in mixture Poisson model other than 4.3.3. For simplicity, we will focus on the Poisson mixture with two components in this section. We also define a simpler latent variable w_i:

$$y_i \in \begin{cases} Group_1 & \text{if } w_i < 0 \\ Group_2 & \text{otherwise} \end{cases}$$

The likelihood for x_i is: $f_P(y_i|\lambda_1)^{I(w_i<0)} f_P(y_i|\lambda_2)^{I(w_i\geq 0)}$

Probit link function was used. We assume that the latent variable $w_i \sim N(x_i\beta, 1)$. The design matrix X_{n*p} contains a column of 1's for intercept and p-1 covariates. We further assume that $rank(X) = p$.

The joint posterior distribution of the parameters is:

$$f(\lambda, \beta, w|x, y) \propto f(y|\lambda, \beta, w, x)\pi(\lambda, \beta, w, x)$$

$$= f(y|\lambda, w)f(w|x, \beta)\pi(\lambda)\pi(\beta)$$

Gaussian and Gamma priors were adopted:

$$\pi(\beta) = N(\beta_0, \Sigma_0)$$

$$\pi(\lambda_j) = Gamma(a_j, b_j)$$

The full conditional distributions for the parameters are discussed as below:

$$f(w_i|.) \propto f_P(y_i|\lambda_1)^{I(w_i<0)} f_P(y_i|\lambda_2)^{I(w_i\geq 0)} N(x_i\beta, 1)$$

$$= f_P(y_i|\lambda_1)N(x_i\beta, 1)I(w_i < 0) + f_P(y_i|\lambda_2)N(x_i\beta, 1)I(w_i \geq 0)$$

$f(w_i|.)$ is a mixture of truncated normal.

$$f(\beta|.) \propto (\prod_{i=1}^{n} N(x_i\beta, 1))N(\beta_0, \Sigma_0)$$

$$\Rightarrow f(\beta|.) = N((\Sigma_0^{-1} + X^T X)^{-1}(\Sigma_0^{-1}\beta_0 + X^T w), (\Sigma_0^{-1} + X^T X)^{-1})$$

$f(\beta|.)$ is a normal distribution.

$$f(\lambda_1|.) \propto (\prod_{i=1}^{n} f_P(y_i|\lambda_1)^{I(w_i<0)})Gamma(a_1, b_1)$$

$$\Rightarrow f(\lambda_1|.) = Gamma(a_1 + \sum_{i=1}^{n} y_i I(w_i < 0), b_1 + \sum_{i=1}^{n} I(w_i < 0))$$

$$f(\lambda_2|.) \propto (\prod_{i=1}^{n} f_P(y_i|\lambda_2)^{I(w_i\geq 0)})Gamma(a_2, b_2)$$

$$\Rightarrow f(\lambda_2|.) = Gamma(a_2 + \sum_{i=1}^{n} y_i I(w_i \geq 0), b_2 + \sum_{i=1}^{n} I(w_i \geq 0))$$

$f(\lambda_j|.)$ is a gamma distribution.

The sampling procedure for $f(w_i|.)$is as follows:

1. Calculate $p = \frac{f_P(y_i|\lambda_2)}{f_P(y_i|\lambda_1)+f_P(y_i|\lambda_2)}$
2. Generate $u \sim Bernoulli(p)$
3. If $u = 0$, then generate $w_i \sim N(x_i\beta, 1)$ truncated at 0 by the right; Otherwise, generate $w_i \sim N(x_i\beta, 1)$ truncated at 0 by the left.

4.3.5 Example: CEI Data

We applied the Bayesian Poisson mixture model with regression on proportion to a traffic accident data set, which contains the crash history and demographic information for more than three thousand drivers. The response is the number of crashes in consecutive seven years, and the related variables include:

- Age
- Gender (1: male; 0: female)
- Role (1: manager; 0: otherwise)
- Motor Vehicle Record (MVR) count

Table 4.1 shows the summary of posterior samples:

Table 4.1: Parameter Estimation for CEI Data: Poisson Mixture

Name	Posterior Mean	95% C.I. Lower Bound	95% C.I. Upper Bound
λ_1	0.932	0.86	0.99
λ_2	2.72	2.64	2.82
$\beta_0(Intercept)$	3.94	1.92	6.02
$\beta_1(Age)$	-0.088	-0.13	-0.044
$\beta_2(Gender)$	-1.1	-1.74	-0.44
$\beta_3(Role)$	-1.22	-2.15	-0.30
$\beta_4(MVR)$	1.44	1.04	1.85

All the estimates are significant. As the age increases, the driver will be more likely to be in the lower-risk group. Male and managers were safer in general. The MVR is also a significant predictor for risky driver.

If the model is applied to identify drivers into safe or risky group, the potential misclassification problem should be taken into account. We will continue this discussion based on the overlap probability.

4.3.6 Negative Binomial VS Poisson

It is well known that Poisson distribution assumes that the mean and variance should be the same, which is too restricted for practical use. The term "overdispersion" is used to

describe the pattern that the variance of the data is significantly larger than the mean. Under overdispersion conditions, negative binomial distribution should be considered as a replacement of Poisson (Gardner et al. 1995).

The probability mass of negative binomial is usually written as:

$$P(X = k|r, p) = \binom{k + r - 1}{k}(1 - p)^r p^k$$

An alternative form is:

$$P(X = k|m, d) = (\frac{d}{d + m})^d * \frac{\Gamma(d + k)}{k!\Gamma(d)} * (\frac{m}{m + d})^k,$$

where m is the mean and d is the dispersion parameter. The variance is $m + \frac{m^2}{d}$. When d goes to infinity, the negative binomial will degenerate to Poisson. Therefore, it is possible to test the hypothesis that $H_0 : d = 0$ to see if negative binomial is necessary.

Dean and Lawless (1989) proposed a simple way to test if Poisson regression is good enough for the data versus negative binomial. It is well known that the variance of negative binomial is a quadratic form of the mean. Assuming that

$$var(X) = \mu + \tau * \mu^2$$

We want to test that:

$$H_0 : \tau = 0; H_1 : \tau > 0$$

The test statistic is:

$$P = \sum_{i=1}^{n} (X_i - \hat{\mu_i}^2)/\bar{X}$$

$\hat{\mu_i}$ is the MLE of mean parameter. Under the null hypothesis, if n goes to infinity then $P/\sqrt{0.5 * \sum \mu_i^2}$ converges to standard normal. If the sample size is not large enough to use normal approximation, then a simulation (as Bootstrap) based empirical distribution is needed.

4.4 Overlap Probability

4.4.1 Overlap Probability in Two Poisson Distributions

For two Poisson distributions, the overlap probability, denoted by $OP_{P,P}(\lambda_1, \lambda_2)$ is defined as:

Definition 4.4.1.

$$OP_{P,P}(\lambda_1, \lambda_2) = \sum_{i=0}^{+\infty} min(f_P(i|\lambda_1), f_P(i|\lambda_2))$$

The following lemma provides an alternative way to calculate the overlap probability.

Lemma 4.4.1. *Assume $\lambda_2 > \lambda_1$, and they are both positive real numbers,*

$$OP_{P,P}(\lambda_1, \lambda_2) = \sum_{k=0}^{[\frac{\lambda_2-\lambda_1}{log(\lambda_2/\lambda_1)}]} \frac{\lambda_2^k e^{-\lambda_2}}{k!} + \sum_{k=[\frac{\lambda_2-\lambda_1}{log(\lambda_2/\lambda_1)}]+1}^{+\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!}$$

Proof. The condition for k when $f_P(k|\lambda_1)$ is greater than $f_P(k|\lambda_2)$ is:

$$f_P(k|\lambda_1) > f_P(k|\lambda_2)$$

$$\Leftrightarrow \lambda_1^k e^{-\lambda_1} > \lambda_2^k e^{-\lambda_2}$$

$$\Leftrightarrow k < \frac{\lambda_2 - \lambda_1}{log(\lambda_2/\lambda_1)}$$

In general, $\frac{\lambda_2-\lambda_1}{log(\lambda_2/\lambda_1)}$ is not an integer. Thus, we may use its integer part in the lower and upper bounds of summation.

□

If the parameters of two Poisson distributions are far from each other, then intuitively their overlap probability should be small. Theorem 4.4.1 confirms that OP is a valid measurement for the magnitude of overlap.

Lemma 4.4.2. *If $x > c > 0$, then $x > \frac{x-c}{log(x)-log(c)}$*

Proof. Define $g(x) = x(log(x) - log(c)) - (x - c)$. It is easy to show that:

$$g(c) = 0$$

$$\frac{dg(x)}{dx} = log(x) - log(c)$$

Because $x > c > 1$, we have:

$$g(x) > 0$$

$$\Rightarrow x > \frac{x - c}{log(x) - log(c)}$$

□

Corollary 4.4.1. *For $x > c > 0$, $h(x) = \frac{x-c}{log(x)-log(c)}$ is a monotone increasing function with respect to x.*

Proof. It is easy to show that:

$$\frac{dh(x)}{dx} = \frac{x(log(x) - log(c)) - (x - c)}{x(log(x) - log(c))^2}$$

According to Lemma 4.4.2, $\frac{\partial h(x)}{\partial x} > 0$.

□

Lemma 4.4.3. *For $\lambda > 0$, $k \in N$, $I(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ is monotone decreasing when $\lambda > k$*

Proof.

$$\frac{\partial I(\lambda)}{\partial \lambda} = \frac{\lambda^k e^{-\lambda}(k/\lambda - 1)}{k!} < 0 \text{ if } \lambda > k$$

□

Theorem 4.4.1. *Assume $\lambda_2 > \lambda_1$, and they are both positive real numbers. For fixed λ_1, $OP_{P,P}(\lambda_1, \lambda_2)$ is a monotone decreasing function with respect to λ_2.*

Proof. Assume λ_2^* is a positive real number s.t. $\lambda_2^* > \lambda_2$, we need to show that $OP_{P,P}(\lambda_1, \lambda_2) > OP_{P,P}(\lambda_1, \lambda_2^*)$. By Corollary 4.4.1, $[\frac{\lambda_2 - \lambda_1}{log(\lambda_2/\lambda_1)}] \leq [\frac{\lambda_2^* - \lambda_1}{log(\lambda_2^*/\lambda_1)}]$. There are two possible situations:

- $[\frac{\lambda_2 - \lambda_1}{log(\lambda_2/\lambda_1)}] = [\frac{\lambda_2^* - \lambda_1}{log(\lambda_2^*/\lambda_1)}]$
- $[\frac{\lambda_2 - \lambda_1}{log(\lambda_2/\lambda_1)}] + T = [\frac{\lambda_2^* - \lambda_1}{log(\lambda_2^*/\lambda_1)}]$, T is some positive integer

For the first situation According to Lemma 4.4.2, $[\frac{\lambda_2 - \lambda_1}{log(\lambda_2/\lambda_1)}] < \lambda_2$, thus by Lemma 4.4.3 under the condition $\lambda_2^* > \lambda_2$ we have:

$$\sum_{k=0}^{[\frac{\lambda_2^* - \lambda_1}{log(\lambda_2^*/\lambda_1)}]} \frac{\lambda_2^{*k} e^{-\lambda_2^*}}{k!} < \sum_{k=0}^{[\frac{\lambda_2 - \lambda_1}{log(\lambda_2/\lambda_1)}]} \frac{\lambda_2^k e^{-\lambda_2}}{k!}$$

Therefore, $OP_{P,P}(\lambda_1, \lambda_2) > OP_{P,P}(\lambda_1, \lambda_2^*)$.

For the second situation

$$[\frac{\lambda_2 - \lambda_1}{log(\lambda_2/\lambda_1)}] + T = [\frac{\lambda_2^* - \lambda_1}{log(\lambda_2^*/\lambda_1)}]$$

Denote $[\frac{\lambda_2^*-\lambda_1}{log(\lambda_2^*/\lambda_1)}]$ as S. The summation can be split into four parts:

$$\begin{aligned}
&OP_{P,P}(\lambda_1, \lambda_2^*) \\
&= \sum_{k=0}^{S} \frac{\lambda_2^{*k} e^{-\lambda_2^*}}{k!} + \sum_{k=S+1}^{S+T} \frac{\lambda_2^{*k} e^{-\lambda_2^*}}{k!} - \sum_{k=S+1}^{S+T} \frac{\lambda_1^k e^{-\lambda_1}}{k!} + \sum_{k=S+1}^{+\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \\
&< OP_{P,P}(\lambda_1, \lambda_2) + \sum_{k=S+1}^{S+T} \frac{\lambda_2^{*k} e^{-\lambda_2^*}}{k!} - \sum_{k=S+1}^{S+T} \frac{\lambda_1^k e^{-\lambda_1}}{k!}
\end{aligned}$$

By Lemma 4.4.2, $S+T = [\frac{\lambda_2^*-\lambda_1}{log(\lambda_2^*/\lambda_1)}] < \lambda_2^*$. As the same logic in the first situation,

$$\Rightarrow \sum_{k=S+1}^{S+T} \frac{\lambda_2^{*k} e^{-\lambda_2^*}}{k!} - \sum_{k=S+1}^{S+T} \frac{\lambda_1^k e^{-\lambda_1}}{k!} < 0$$

Thus we have:

$$OP_{P,P}(\lambda_1, \lambda_2^*) < OP_{P,P}(\lambda_1, \lambda_2)$$

Therefore, the inequality $OP_{P,P}(\lambda_1, \lambda_2^*) < OP_{P,P}(\lambda_1, \lambda_2)$ holds for both situations. □

4.4.2 OP as a Metric

The overlap probability can be used to define a metric in the space of Poisson probability mass functions.

Theorem 4.4.2. *$d = 1 - OP_{P,P}$ is a metric defined for the Poisson probability mass functions.*

Proof. We have to verify four properties.

(1) $d \geq 0$

(2) $d(f_1, f_2) = 0$ if and only if $f_1 = f_2$

(3) $d(f_1, f_2) = d(f_2, f_1)$

(4) $d(f_1, f_2) + d(f_2, f_3) \geq d(f_1, f_3)$

(1) and (3) are obvious. For (2), $f_1 = f_2 \Rightarrow d(f_1, f_2) = 0$ is also trivial.

We will prove $d(f_1, f_2) = 0 \Rightarrow f_1 = f_2$.

$$d(f_1, f_2) = 0 \Leftrightarrow OP_{P,P}(\lambda_1, \lambda_2) = 1$$

$$\Rightarrow \sum_{i=0}^{+\infty} min(f_P(i|\lambda_1), f_P(i|\lambda_2)) = 1 = \sum_{i=0}^{+\infty} f_P(i|\lambda_1)$$

It is known that $\forall i, min(f_P(i|\lambda_1), f_P(i|\lambda_2)) \leq f_P(i|\lambda_1)$

If $\exists j, min(f_P(j|\lambda_1), f_P(j|\lambda_2)) < f_P(j|\lambda_1)$,

then $\sum_{i=0}^{+\infty} min(f_P(i|\lambda_1), f_P(i|\lambda_2)) < \sum_{i=0}^{+\infty} f_P(i|\lambda_1)$, which leads to contradiction.

Therefore, $\forall i, min(f_P(i|\lambda_1), f_P(i|\lambda_2)) = f_P(i|\lambda_1)$

Similarly, $\forall i, min(f_P(i|\lambda_1), f_P(i|\lambda_2)) = f_P(i|\lambda_2)$

$$\Rightarrow \lambda_1 = \lambda_2$$

(4) is triangle inequality. Without loss of generality, we only need to verify two cases: $\lambda_1 < \lambda_2 < \lambda_3$ and $\lambda_1 < \lambda_3 < \lambda_2$. For the second case, according to Theorem 4.4.1 it is trivial. Therefore, we will work on the first case. Under such condition,

$$d(f_1, f_2) + d(f_2, f_3) \geq d(f_1, f_3) \Leftrightarrow 1 + OP_{P,P}(\lambda_1, \lambda_3) \geq OP_{P,P}(\lambda_1, \lambda_2) + OP_{P,P}(\lambda_2, \lambda_3)$$

A detailed derivation could be slightly tedious. The Figure 4.4 might be more intuitive (the

mass functions have been made continuous to be easy to see).

$$1 + OP_{P,P}(\lambda_1, \lambda_3) = 1 + A3$$

$$OP_{P,P}(\lambda_1, \lambda_2) = A1 + A3$$

$$OP_{P,P}(\lambda_2, \lambda_3) = A3 + A4$$

$$1 + OP_{P,P}(\lambda_1, \lambda_3) - OP_{P,P}(\lambda_1, \lambda_2) - OP_{P,P}(\lambda_2, \lambda_3)$$

$$= (1 + A3) - (A1 + A3 + A3 + A4)$$

$$= (A1 + A2 + A3 + A4 + A3) - (A1 + A3 + A3 + A4)$$

$$= A2 \geq 0$$

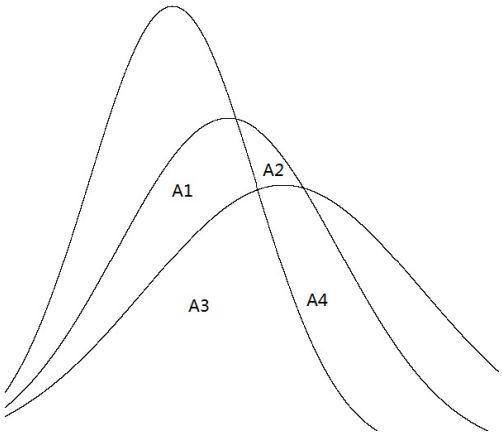

Figure 4.4: Illustration of Triangle Inequality

□

This metric can be used for hierarchical clustering (Johnson 1967). A hierarchical clustering model generally requires two tuning parameters: the metric and the linkage criteria, which define the way to calculate the distance between two elements and two sets respectively. In

this paper we stick to the complete-linkage, which is generally the distance between those two elements (one in each cluster) that are farthest away from each other and avoids a drawback of the single linkage method where clustering may be forced together due to a single pair of elements.

The choice of metric could also influence the shape of the clusters since some elements may be close to one another in one distance and farther away according to another. The most common metric is the Euclidean metric. Suppose there are four drivers with annual accident rates: 0.1, 1.1, 8, 9, the Euclidean metric claims that the distance between 0.1 and 1.1 is the same as the distance from 8 to 9. However, the intuition tells us that to distinguish the first pair is much easier because the second pair has significantly larger variance.

As an example, we will show that the hierarchical clustering result by Euclidean metric differs from that by OP metric. We generate 10 ascending values (#1-#10) for Poisson parameter λ: 1.81, 1.83, 2.59, 4.49, 5.8, 6.98, 7.60, 8.64, 9.19, 9.51 and the dendrograms are as follows.

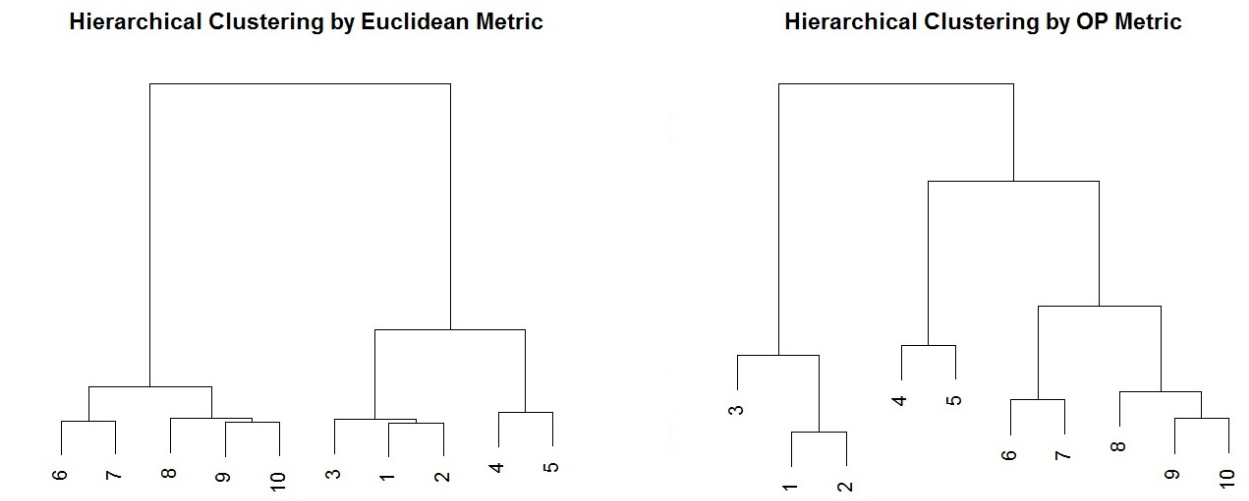

Figure 4.5: Comparison of Clustering Results

The dendrograms in Figure 4.5 indicate the clear difference in clustering results by different metrics. If the drivers are to be separated into two groups, the Euclidean result is {#1 to #5} {#6 to #10} while the OP result shows {#1 to #3} {#4 to #10}.

In order to validate the clustering result, we have set up a simple simulation study.

Suppose $X \sim 0.5 * Poisson(\lambda_1) + 0.5 * Poisson(\lambda_2)$, λ_i is randomly sampled from cluster i (i=1 or 2). We generate two data sets by the two different clustering structure. For each data set we use EM algorithm to predict whether an observation is safe or risky. The classification results are:

Table 4.2: Classification Results Comparison

	OP		Euclidean	
	Pred. Safe	Pred. Risky	Pred. Safe	Pred. Risky
True Safe	455	45	350	150
True Risky	98	402	53	447
Misclassification Rate	0.143		0.203	

Table 4.2 shows that the clustering structure obtained by OP metric can be a better predictor to identify safe and riky drivers.

One of the popular measure of the difference between two probability distributions is Kullback-Leibler divergence (Kullback and Leibler 1951). Given two distributions P and Q, it is defined as:

$$D_{KL}(P||Q) = \sum_i ln(\frac{P(i)}{Q(i)})P(i)$$

It is easy to show that suppose P and Q are two Poisson distributions with parameters λ_2 and λ_1, the Kullback-Leibler divergence is: $\lambda_2 * ln(\lambda_2/\lambda_1) - (\lambda_2 - \lambda_1)$. It is monotonic increasing with λ_2 when λ_1 is fixed. However, Kullback-Leibler divergence is not symmetric so can not be defined as a metric.

4.4.3 Overlap Probability in Mixture Possion

In this section, we will generalize the idea of overlap probability from two Poisson distributions to a two-component Poisson distribution.

Definition 4.4.2.

$$OP_{P(2)}(\lambda_1, \lambda_2, \alpha, C) = \sum_{k=0}^{[C]}(1-\alpha)\frac{\lambda_2^k e^{-\lambda_2}}{k!} + \sum_{k=[C]+1}^{+\infty}\alpha\frac{\lambda_1^k e^{-\lambda_1}}{k!},$$

where $C \geq 0$.

The above definition of the overlap probability for a two-component Poisson distribution has a similar form as Definition 4.4.1. In fact, when $\alpha = 0.5$ the $OP_{P(2)}$ is exactly one half of $OP_{(P,P)}$.

Definition 4.4.3. *The* ***simple decision rule*** *can be stated as: If* $\pi_{i2} > \pi_{i1}$ *(which is equivalent to* $\pi_{i2} > 0.5$*), then the individual* x_i *will be assigned to Group 2 and vice versa.*

The overlap probability can be used to estimate the misclassification rate for EM algorithm under the simple decision rule. Suppose the EM algorithm yields the estimates as $\hat{\lambda_1}, \hat{\lambda_2}$ and $\hat{\alpha}$, it is easy to see that the probability that the observations from Group 1 is classified into Group 2 is:

$$\sum_{k=[\hat{C}]+1}^{+\infty}\alpha\frac{\lambda_1^k e^{-\lambda_1}}{k!},$$

and the probability that the observations from Group 2 is classified into Group 1 is:

$$\sum_{k=0}^{[\hat{C}]}(1-\alpha)\frac{\lambda_2^k e^{-\lambda_2}}{k!},$$

where

$$\hat{C} = \frac{\hat{\lambda_2} - \hat{\lambda_1} + log(\hat{\alpha}/(1-\hat{\alpha}))}{log(\hat{\lambda_2}/\hat{\lambda_1})} \geq 0.$$

The proof is directly from the definition of simple decision rule. The sum of the above two quantities is exactly the overlap probability. When $\hat{C} < 0$ (i.e. $\hat{\alpha} < (1 + e^{\hat{\lambda_2} - \hat{\lambda_1}})^{-1}$), all the observations will be classified as in group 2 and in such case it is reasonable to define the overlap probability as α.

The estimation from EM algorithm could hardly be the exact true values. Actually, the true values of the parameters provide an "lower bound" of the estimated misclassification rate:

Lemma 4.4.4. *As a function of C, $OP_{P(2)}(\lambda_1, \lambda_2, \alpha, C)$ reaches its minimum value when*

$$C = \frac{\lambda_2 - \lambda_1 + log(\alpha/(1-\alpha))}{log(\lambda_2/\lambda_1)} \geq 0 \text{ (We may denote this value as } C')$$

Proof. Solve the equation for x:

$$\alpha \frac{e^{-\lambda_1}\lambda_1^x}{x!} = (1-\alpha)\frac{e^{-\lambda_2}\lambda_2^x}{x!}$$

□

Usually the true values of the parameters are unknown in the real-life data sets. We will do a simple simulation study.

We randomly pick 5 combinations of the λ_1, λ_2 and α. For each combination, we simulate 1000 observations from the mixture Poisson distribution and apply EM algorithm to do the clustering. The initial values are randomly selected.

Table 4.3: Simulation Study for Overlap Probability

	1	2	3	4	5
α	0.5	0.8	0.5	0.2	0.5
λ_1	0.2	0.2	0.2	0.2	1
λ_2	0.8	0.8	0.6	0.6	10
Observed miscl. rate	0.435	0.178	0.505	0.470	0.012
Estimated miscl. rate	0.413	0.176	0.500	0.475	0.015
Lower bound	0.315	0.176	0.365	0.200	0.015

From Table 4.3 the estimated misclassification rates are generally close to the observed ones, and they are always greater than or equal to the lower bounds as we expect.

4.4.4 Discussion on $OP_{P(2)}(\lambda_1, \lambda_2, \alpha, C')$

In this section we will focus on $OP_{P(2)}(\lambda_1, \lambda_2, \alpha, C')$. The value of C will be fixed at C' hereafter. A natural question is whether it also follows the monotonicity similar as Theorem 4.4.1. Firstly, we proved a slightly weaker result, which is the locally monotonicity.

Lemma 4.4.5. *Assume $\lambda_2 > \lambda_1 > 0$, $C' \geq 0$, and α is a real number within (0,1). For fixed λ_1, $OP_{P(2)}(\lambda_1, \lambda_2, \alpha)$ is a locally monotone decreasing function with respect to λ_2:*

For a small positive real number h, s.t.

$$[\frac{\lambda_2 - \lambda_1 + log(\alpha/(1-\alpha))}{log(\lambda_2/\lambda_1)}] = [\frac{\lambda_2 + h - \lambda_1 + log(\alpha/(1-\alpha))}{log((\lambda_2+h)/\lambda_1)}]$$

We have:

$$OP_{P(2)}(\lambda_1, \lambda_2, \alpha) > OP_{P(2)}(\lambda_1, \lambda_2 + h, \alpha)$$

Proof. Since the offset h will not change $[C']$, it is easy to see that we only need to prove:

$$\sum_{k=0}^{[C']} \frac{\lambda_2^k e^{-\lambda_2}}{k!} > \sum_{k=0}^{[C']} \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!}$$

By Lemma 4.4.3, we claim that as long as $[C'] \leq \lambda_2$, the inequality must hold. From now on, we will assume that $[C'] > \lambda_2$. By Lemma 4.4.2, this implies that $\alpha > 0.5$.

Let's consider two possible situations:

- $[C'] \geq \lambda_2 + h$
- $\lambda_2 < [C'] < \lambda_2 + h$

For the first situation

Our objective is to prove that:

$$\sum_{k=0}^{[\lambda_2]} \frac{(\lambda_2)^k e^{-\lambda_2}}{k!} - \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!} > \sum_{k=[\lambda_2]+1}^{[C']} \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!} - \frac{(\lambda_2)^k e^{-\lambda_2}}{k!}$$

It is easy to see that:

$$\begin{aligned}
&\sum_{k=[\lambda_2]+1}^{[C']} \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!} - \frac{(\lambda_2)^k e^{-\lambda_2}}{k!} < \sum_{k=[\lambda_2]+1}^{+\infty} \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!} - \frac{(\lambda_2)^k e^{-\lambda_2}}{k!} \\
&= (1 - \sum_{k=0}^{[\lambda_2]} \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!}) - (1 - \sum_{k=0}^{[\lambda_2]} \frac{(\lambda_2)^k e^{-\lambda_2}}{k!}) \\
&= \sum_{k=0}^{[\lambda_2]} \frac{(\lambda_2)^k e^{-\lambda_2}}{k!} - \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!}
\end{aligned}$$

For the second situation

By solving the equation of k:

$$\lambda_2^k e^{-\lambda_2} = (\lambda_2 + h)^k e^{-\lambda_2 - h}$$

We get the "imaginary intersection" of two Poisson mass functions: $k = \frac{h}{log(1+h/\lambda_2)}$.

If $[C'] \leq \frac{h}{log(1+h/\lambda_2)}$, then within the range of $\{0, 1, 2, ..., [C']\}$ we will have:

$$\begin{aligned} &\lambda_2^k e^{-\lambda_2} > (\lambda_2 + h)^k e^{-\lambda_2 - h} \\ \Longrightarrow &\sum_{k=0}^{[C']} \frac{\lambda_2^k e^{-\lambda_2}}{k!} > \sum_{k=0}^{[C']} \frac{(\lambda_2 + h)^k e^{-(\lambda_2 + h)}}{k!} \end{aligned}$$

If $[C'] > \frac{h}{log(1+h/\lambda_2)}$, we have:

$$\begin{cases} \lambda_2^k e^{-\lambda_2} > (\lambda_2 + h)^k e^{-\lambda_2 - h} & \text{, if } k \leq [\frac{h}{log(1+h/\lambda_2)}] \\ \lambda_2^k e^{-\lambda_2} < (\lambda_2 + h)^k e^{-\lambda_2 - h} & \text{, if } k > [\frac{h}{log(1+h/\lambda_2)}] \end{cases}$$

By Lemma 4.4.3:

$$\sum_{k=0}^{[\lambda_2+h]} \frac{\lambda_2^k e^{-\lambda_2}}{k!} > \sum_{k=0}^{[\lambda_2+h]} \frac{(\lambda_2 + h)^k e^{-(\lambda_2+h)}}{k!}$$

If $[C'] = [\lambda_2 + h]$ then the proof has finished. If $[C'] < [\lambda_2 + h]$:

$$\begin{aligned} &\Longrightarrow \sum_{k=[C']+1}^{[\lambda_2+h]} \frac{\lambda_2^k e^{-\lambda_2}}{k!} < \sum_{k=[C']+1}^{[\lambda_2+h]} \frac{(\lambda_2 + h)^k e^{-(\lambda_2+h)}}{k!} \\ &\Longrightarrow \sum_{k=0}^{[\lambda_2+h]} \frac{\lambda_2^k e^{-\lambda_2}}{k!} - \sum_{k=[C']+1}^{[\lambda_2+h]} \frac{\lambda_2^k e^{-\lambda_2}}{k!} > \sum_{k=0}^{[\lambda_2+h]} \frac{(\lambda_2 + h)^k e^{-(\lambda_2+h)}}{k!} - \sum_{k=[C']+1}^{[\lambda_2+h]} \frac{(\lambda_2 + h)^k e^{-(\lambda_2+h)}}{k!} \\ &\Longrightarrow \sum_{k=0}^{[C']} \frac{\lambda_2^k e^{-\lambda_2}}{k!} > \sum_{k=0}^{[C']} \frac{(\lambda_2 + h)^k e^{-(\lambda_2+h)}}{k!} \end{aligned}$$

□

It is worth noting that because of the impact from $log(\alpha/(1-\alpha))$, C' is not always increasing when λ_2 increases.

$$\frac{\partial C'}{\partial \lambda_2} = \frac{\lambda_2 log(\lambda_2/\lambda_1) - (\lambda_2 - \lambda_1 + log(\alpha/(1-\alpha)))}{\lambda_2 (log(\lambda_2/\lambda_1))^2}$$

If α is large enough, when λ_2 is increasing, the position of C' might be shifted to the left.

Lemma 4.4.6. *Assume that*

$$0 \leq [C'] = [\frac{\lambda_2 - \lambda_1 + log(\alpha/(1-\alpha))}{log(\lambda_2/\lambda_1)}] = [\frac{\lambda_2 + h - \lambda_1 + log(\alpha/(1-\alpha))}{log((\lambda_2+h)/\lambda_1)}] + 1$$

We have:

$$OP_{P(2)}(\lambda_1, \lambda_2, \alpha) > OP_{P(2)}(\lambda_1, \lambda_2 + h, \alpha)$$

Proof. By definition,

$$\begin{aligned} &OP_{P(2)}(\lambda_1, \lambda_2, \alpha) - OP_{P(2)}(\lambda_1, \lambda_2 + h, \alpha) \\ &= (1-\alpha)\frac{(\lambda_2+h)^{[C']} e^{-(\lambda_2+h)}}{[C']!} - \alpha \frac{(\lambda_1)^{[C']} e^{-\lambda_1}}{[C']!} + \sum_{k=0}^{[C']} (\frac{\lambda_2^k e^{-\lambda_2}}{k!} - \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!}) \end{aligned}$$

By similar derivation as in Lemma 4.4.5, we conclude that $\sum_{k=0}^{[C']} (\frac{\lambda_2^k e^{-\lambda_2}}{k!} - \frac{(\lambda_2+h)^k e^{-(\lambda_2+h)}}{k!}) > 0$.

So it is sufficient to show that:

$$\begin{aligned} &(1-\alpha)\frac{(\lambda_2+h)^{[C']} e^{-(\lambda_2+h)}}{[C']!} > \alpha \frac{(\lambda_1)^{[C']} e^{-\lambda_1}}{[C']!} \\ \iff &log(\frac{\alpha}{1-\alpha}) < [C'] log(\frac{\lambda_2+h}{\lambda_1}) - (\lambda_2 + h - \lambda_1) \\ \iff &\frac{log(\frac{\alpha}{1-\alpha}) + (\lambda_2 + h - \lambda_1)}{log(\frac{\lambda_2+h}{\lambda_1})} < [C'] \end{aligned}$$

The last inequality holds according to our assumption. □

By the same logic, we conclude that:

Lemma 4.4.7. *Assume that*

$$0 \leq [C'] = [\frac{\lambda_2 - \lambda_1 + log(\alpha/(1-\alpha))}{log(\lambda_2/\lambda_1)}] = [\frac{\lambda_2 + h - \lambda_1 + log(\alpha/(1-\alpha))}{log((\lambda_2+h)/\lambda_1)}] - 1$$

We have:

$$OP_{P(2)}(\lambda_1, \lambda_2, \alpha) > OP_{P(2)}(\lambda_1, \lambda_2 + h, \alpha)$$

If we combine Lemma 4.4.5, 4.4.6 and 4.4.7, and by the mathematical induction, we will have the following result:

Theorem 4.4.3. *Assume $\lambda_2 > \lambda_1 > 0$, $C' \geq 0$, and α is a real number within (0,1). For fixed λ_1, $OP_{P(2)}(\lambda_1, \lambda_2, \alpha)$ is a monotone decreasing function with respect to λ_2.*

The next question is: How does the values of α affect the overlap probability? First let's look at the relationship between α and C'. It can be shown that C' is monotone increasing with respect to α if $\alpha \in (0, 1)$. The reason is simple:

$$\frac{\partial C'}{\partial \alpha} = \frac{\frac{1}{\alpha} + \frac{1}{1-\alpha}}{log(\lambda_2/\lambda_1)}$$

Thus when α increases, the $[C']$ will either stay the same or move to the right.

When we look at the definition of overlap probability, it is clear that it consists of two parts.

$$(1-\alpha) * S_1 + \alpha * S_2 = \sum_{k=0}^{[C']} (1-\alpha) \frac{\lambda_2^k e^{-\lambda_2}}{k!} + \sum_{k=[C']+1}^{+\infty} \alpha \frac{\lambda_1^k e^{-\lambda_1}}{k!}$$

If the value of α increase, $1-\alpha$ will decrease. By the above lemma, $[C']$ might increase and thus S_1 might increase while S_2 might decrease. Thus the relationship between α and OP is not obvious. We might want to know: given the value of $\lambda_i' s$, can we find an α such that

make the overlap probability reaches its maximum or minimum values? Before answering this question, we have to show that the minimum or maximum values do exist.

Theorem 4.4.4. *Fix the values of λ_1, λ_2 and consider OP as a function of α. We also define $OP(0) = OP(1) = 0$, then it is a continuous function for $\alpha \in [0, 1]$*

Proof. First we prove that:

$$\lim_{\alpha \to 0^+} OP(\alpha) = 0$$

$log(\alpha/(1-\alpha))$ is a monotone increasing function defined within $(0, 1)$ and when α goes to 0 the function will go to $-\infty$. Therefore C' will become negative when $\alpha \to 0^+$. By the definition of OP, under such case the overlap probability will be defined as α so it will also go to 0.

On the other hand,

$$\lim_{\alpha \to 1^-} OP(\alpha) = 0$$

When α goes to 1, C' will go to $+\infty$, and hence $\alpha * S_2$ will go to 0. Meanwhile, $(1-\alpha) * S_1$ also goes to 0. Therefore, OP will go to 0.

So we have proved that for the two boundary points the function is continuous from one direction. When α is so small that $C' < 0$, the proof is also simple. Thus we assume $C' \geq 0$ from now on. There are two situations:

- $C'(\alpha)$ is not an integer
- $C'(\alpha)$ is an integer

For the first situation

For $\forall \epsilon > 0$, first we can pick some $\delta_0 > 0$ such that $[C'(\alpha \pm \delta_0)] = [C'(\alpha)]$ since the integer part is a step function and $C'(\alpha)$ is continuous. So we have:

$$OP(\alpha) = (1-\alpha) * S_1 + \alpha * S_2$$

$$OP(\alpha \pm \delta) = (1 - (\alpha \pm \delta_0)) * S_1 + (\alpha \pm \delta_0) * S_2$$

$$\Longrightarrow |OP(\alpha \pm \delta) - OP(\alpha)| = \delta_0 |S_1 - S_2|$$

Assume that $|S_1 - S_2| > 0$, if we select a number $\delta = \frac{1}{2} * min(\delta_0, \frac{\epsilon}{|S_1 - S_2|})$, then as long as $|\alpha^* - \alpha| < \delta$, we have $|OP(\alpha^*) - OP(\alpha)| < \epsilon$.

For the second situation

For $\forall \epsilon > 0$, it is easy to see that we can still pick some $\delta_0 > 0$ such that $[C'(\alpha + \delta_0)] = [C'(\alpha)]$ and do the similar procedure. So next we will consider that $[C'(\alpha - \delta_0)] = [C'(\alpha)] - 1$.

$$|OP(\alpha - \delta_0) - OP(\alpha)|$$
$$= -\delta_0 \sum_{k=0}^{[C']-1} \frac{\lambda_2^k e^{-\lambda_2}}{k!} + \delta_0 \sum_{k=[C']+1}^{+\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} + (1-\alpha)\frac{\lambda_2^{C'} e^{-\lambda_2}}{C'!} - (\alpha - \delta_0)\frac{\lambda_1^{C'} e^{-\lambda_1}}{C'!}$$

By the definition of C', we know that:

$$(1-\alpha)\frac{\lambda_2^{C'} e^{-\lambda_2}}{C'!} - \alpha \frac{\lambda_1^{C'} e^{-\lambda_1}}{C'!} = 0$$

Thus, $|OP(\alpha - \delta_0) - OP(\alpha)|$

$$= -\delta_0 \sum_{k=0}^{[C']-1} \frac{\lambda_2^k e^{-\lambda_2}}{k!} + \delta_0 \sum_{k=[C']+1}^{+\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!} + \delta_0 \frac{\lambda_1^{C'} e^{-\lambda_1}}{C'!}$$

When we pick δ_0 small enough, e.g. $\epsilon/3$, then $|OP(\alpha^*) - OP(\alpha)| < \epsilon$. □

By calculus we know that if a real-valued function f is continuous in the closed and bounded interval [a,b], then f must attain its maximum and minimum value. Therefore, we claim

that:

Lemma 4.4.8. *The overlap probability as a function of α does attain its maximum and minimum value within [0,1].*

It is easy to see that the minimum value of $OP(\alpha)$ is 0 and is attained at $\alpha = 0$ and $\alpha = 1$. However, to obtain the α which maximizes the overlap probability is not a simple question because the α plays a role in the upper and lower bounds of the summation. Also, a continuous function may not be a differentiable function so it is unclear that if we can get the maximum value by setting $f' = 0$.

A simple way is to use brute force. For example, we set up a number of grids within (0,1) and pick the one with highest OP value. Could we "shrink" our search range so to improve the efficiency? The next theorem will give some hints.

Theorem 4.4.5. *Suppose α^* is the point that $OP(\alpha)$ reaches its local maximum for any of the following three cases:*

- $OP(\alpha^*-) \leq OP(\alpha^*)$ *and* $OP(\alpha^*+) < OP(\alpha^*)$
- $OP(\alpha^*-) < OP(\alpha^*)$ *and* $OP(\alpha^*+) \leq OP(\alpha^*)$
- $OP(\alpha^*-) < OP(\alpha^*)$ *and* $OP(\alpha^*+) < OP(\alpha^*)$

Then α must be some real number that makes $C'(\alpha)$ to be an integer.

Proof. If $C'(\alpha)$ is not an integer, then we can always select an ϵ so small that: $[C'(\alpha^* + \epsilon)] =$

$[C'(\alpha^*)] = [C'(\alpha^* - \epsilon)]$. For situation 1, by definition of OP, we have:

$$(1 - \alpha^* + \epsilon) * S_1 + (\alpha^* - \epsilon) * S_2 \leq (1 - \alpha^*) * S_1 + (\alpha^*) * S_2$$

$$(1 - \alpha^* - \epsilon) * S_1 + (\alpha^* + \epsilon) * S_2 < (1 - \alpha^*) * S_1 + (\alpha^*) * S_2$$

$$\Longrightarrow S_1 \leq S_2, S_2 > S_1$$

This is definitely impossible. Similarly:

- For situation 2, $S_1 < S_2, S_2 \geq S_1$
- For situation 3, $S_1 < S_2, S_2 > S_1$

All of them are impossible. Therefore, the assumption leads to controversy. □

The reader must have noticed that there might be the situation 4: $OP(\alpha^*-) \leq OP(\alpha^*)$ and $OP(\alpha^*+) \leq OP(\alpha^*)$. That is, the OP as a function of α will have a "flat" segment. However, this is only possible when $S_1 = S_2$, i.e.

$$\sum_{k=0}^{[C']} \frac{\lambda_2^k e^{-\lambda_2}}{k!} = \sum_{k=[C']+1}^{+\infty} \frac{\lambda_1^k e^{-\lambda_1}}{k!}$$

The above equality will only hold when λ_1 and λ_2 are the solutions for this particular equation. For a general pair of λ_1 and λ_2, it is highly unlikely that they satisfy the equation by chance.

To sum up, a comprehensive way to find the maximum of $OP(\alpha)$ is as follows:

- Step 1: Set up a proper subset of nonnegative integers, such as $A = \{0, 1, 2, 3, 4, ...m - 1\}$ and solve the corresponding equation for α:

$$\frac{\lambda_2 - \lambda_1 + log(\alpha/(1-\alpha))}{log(\lambda_2/\lambda_1)} = i, \forall i \in A$$

Suppose we denote the m roots as $\{\alpha_1, \alpha_2, ...\alpha_m\}$

- Step 2: Collect all the $\alpha_i's$ and calculate the corresponding overlap probabilities to find the maximum.
- Step 3: If more than one $\alpha_i's$ can attain the same maximum, then all the real numbers within $[min(\alpha_i), max(\alpha_i)]$ should be the solution.

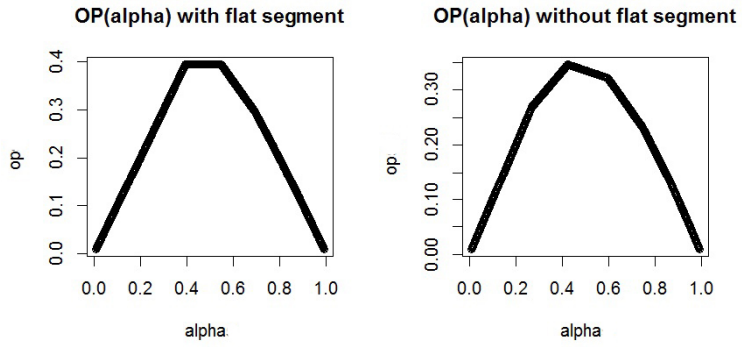

Figure 4.6: Comparison for Two Different Cases in OP(α)

By the way, it is not difficult to see that $OP(\alpha)$ is a concave function.

4.4.5 Example: CEI Data

We fitted the CEI data to illustrate the application of the results in Section 4.4.4.

Suppose for a population represented as female non-manager with no MVR records, what is the relationship between age and the magnitude of "separation" between safe and risky groups?

In Section 4.3.5, we have obtained the estimates of λ_1 and λ_2. According to Theorem 4.4.5, C' is:

$$C' = \frac{2.732 - 0.928 + log(\alpha/(1-\alpha))}{log(2.732/0.928)}$$
$$= \frac{1.803 + log(\alpha/(1-\alpha))}{1.079}$$
$$\Longrightarrow \alpha = \frac{exp(1.079C' + 1.803)}{exp(1.079C' + 1.803) + 1}$$

Assuming $C' = 0, 1, 2, ...$, we will be able to find the maximum in overlap probability with respect to α.

It can be shown that the overlap probability will reach its peak at $\alpha = 0.8586151$, which corresponds to the age of 57. Therefore, for female non-manager with no MVR records, the most "difficult" age for clustering is 57 years old.

4.4.6 Threshold: β

The previous section is based on the simple decision rule: If $\pi_{i2} > \pi_{i1}$ (which is equivalent to $\pi_{i2} > 0.5$), then individual x_i will be assigned to Group 2 and vice versa.

A more flexible decision rule can be adjusted by a scale parameter β, and the general form is: If $\pi_{i2} > \beta * \pi_{i1}$, individual x_i will be assigned to group 2 and vice versa. β is ranging within $(0, +\infty)$.

The lower bound for estimated misclassification rate is attained when

$$C'' = \frac{\lambda_2 - \lambda_1 + log(\alpha/(1-\alpha)) + log(\beta)}{log(\lambda_2/\lambda_1)}$$

It is known that, sometimes people give different weights for the two types of misclassification rates. That leads to the definition of loss function. In our case, the loss function is as follow:

Definition 4.4.4.

$$Loss = \begin{cases} M & \textit{if the observation is from Group 1 but we classify it to 2} \\ 1 & \textit{if the observation is from Group 2 but we classify it to 1} \end{cases}$$

It is not difficult to show that the lower bound of estimated average loss could be obtained when $C = C''$ and $\beta = M$.

Another application is in the case that a particular proportion of one group have to be detected. Suppose the safety managers want to capture at least 80% of the risky drivers. We propose the β s.t.

$$\beta = \sup_{\beta} \{\beta \in (0,1) : \sum_{k=[C'']+1}^{+\infty} f_P(k|\hat{\lambda_2}) \geq 0.8\}$$

Theoretically, the supremum must exist and is unique since the set has upper bound. Note that in practical the supremum can be replaced by the maximum of a finite candidate set.

4.5 Summary

Developing reliable statistical models for analyzing individual driver risk is crucial in transportation studies. Several models have been proposed by the researchers. The finite mixture

Poisson model is based on the premise that the individual driver risk is dominated by the underlying characteristics, which can be described as "high risk" and "low risk", and may have some relationship with demographic information and historical accident records.

Two levels of uncertainty are quantitatively assessed in the finite mixture Poisson model. The first level of uncertainty is the proportion of a given driver group; the second level of uncertainty is the variation of individual risk within each group. EM algorithm and MCMC method can be used to estimate the model.

Our work provides a comprehensive discussion regarding the limitation of the finite mixture Poisson model. The difficulty of classification can be measured by a simple and intuitive concept: overlap probability. A lower bound of the misclassification rate has been proposed and simulation data are used to justify it. We proved that the overlap probability has a monotonic relationship when the parameter of the second component is changed. We also provided an easy way to find the maximum of overlap probability when the proportion varies. CEI data have been used to illustrate some of the theoretical results.

Chapter 5

High-Risk Drivers: Quantile Regression

5.1 Introduction

In this chapter we will discuss about the problem of identifying high-risk drvers. The p-quantile of a random variable X is the value q_p such that:

$$q_p = inf\{x : P(X \leq x) \geq p\}$$

If we assume that the cumulative distribution function of X is continuous and strictly monotonic, q_p can be defined as the value which solves:

$$P(X \leq q_p) = p$$

One example of quantile is the median, which takes p=0.5 and can be seen as the middle value of a data set.

The least square estimate corresponds to the conditional expectation $E(y|X = x)$ as $\hat{\beta}$ minimizes the conditional squared loss $E((y - \hat{y})^2|X = x)$. On the other hand, the least absolute deviation estimate is:

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} |y_i - X_i^T \beta|$$

This $\tilde{\beta}$ leads to the conditional median.

The quantile regression can be seen as an extension of the least absolute deviation regression, as the conditional quantiles rather than median are concerned.

A function called "check function" (Yu et al. 2003) is defined as:

$$\rho_p(x) = p * x * I_{[0,+\infty)}(x) - (1 - p) * x * I_{(-\infty,0)}(x)$$

If p=0.5, then $\rho_{0.5}(x) = 0.5 * |x|$, which is a half of the absolute function. When p is not 0.5, the check function implies an asymmetric pattern as Figure 5.1. When p is selected as 0 or 1, extreme regression quantiles were discussed (Portnoy and Jureckova 1999) and it has a close relationship with extreme value distribution theory.

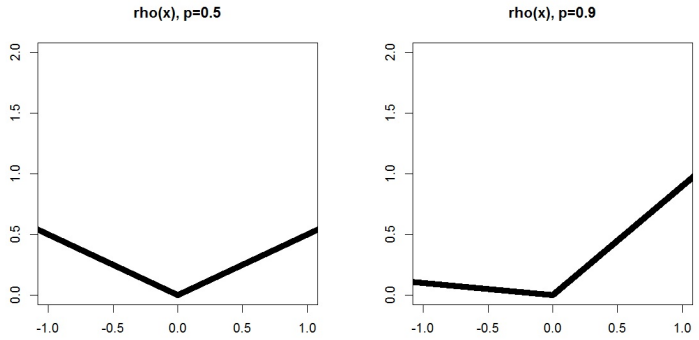

Figure 5.1: Check funtions with different p values

The p-quantile regression obtains the estimate such that:

$$\tilde{\beta}_p = \arg\min_{\beta} \sum_{i=1}^{n} \rho_p(y_i - X_i^T\beta)$$

$$= \arg\min_{\beta} \sum_{i=1}^{n} p * (y_i - X_i^T\beta) * I_{[0,+\infty)}(y_i - X_i^T\beta) - (1-p) * (y_i - X_i^T\beta) * I_{(-\infty,0)}(y_i - X_i^T\beta)$$

It is clear that when p=0.5, the estimate is exactly the least absolute deviation estimate. In general, the p-quantile estimate is corresponding to the conditional p-quantile of y. The proof is simple. In order to minimize:

$$f(u) = p\int_u^{+\infty}(y-u)dF_y - (1-p)\int_{-\infty}^{u}(y-u)dF_y,$$

we take the derivative with respect to u, set the derivative as zero and obtain the root of the equation. According to the basic rule of calculus,

$$\frac{\partial f}{\partial u} = -p\int_u^{+\infty} dF_y + (1-p)\int_{-\infty}^{u} dF_y = 0$$

$$\Longleftrightarrow F(u) - p = 0$$

If we denote the root of the above equation as $\tilde{u}$, then by definition $\tilde{u}$ is the p-quantile.

The quantile regression provides more flexibility compared to standard linear regression. If we are interested in the "average" relationship, a quantile regression with p=0.5 can be applied. Although this result is not the most efficient under the normality assumption, it is more robust to outliers. If p is set to 0.75 or 0.25, we are able to focus on the upper and lower quartiles of the response variable. For example, suppose we have a data set contains the ages and salaries of a group of people and we are more interested in the people with high salaries, a quantile regression with p close to 1 can be used. Similar as ordinary least square regression, quantile regression also has a set of asymptotic properties (Koenker and Bassett 1978). Bootstrap can also be used to establish confidence interval (Jinyong 1995, Koenker 1994).

Consider a common question in safety study: What is the threshold of the top 10% risky drivers in predicted accident rate? Traditionally the researchers construct a standard regression model on accident rates and the answer will be 90th quantile of the fitted values. In quantile regression the 90th conditional quantile for each driver can be directly estimated. The second answer is more logical since the 90th quantile of the mean is not necessarily the mean of the 90th quantile.

One of the reason that people prefer least square in old days is that the square function is easy to differentiate. However, there is no closed-form solution to the regression coefficients estimation because the check function used in quantile regression is not differentiable at the origin. Simplex method (Barrodale and Roberts 1973) and interior point algorithm (Koenker and Park 1996) can be used in the quantile regression problems. MCMC, de-

rived from Bayesian statistics also has its advantages (Yu and Moyeed 2001, Kozumi and Kobayashi 2011, Lancaster and Jae Jun 2010). Koutsourelis and Yu (2012) mentioned that by Bayesian analysis the credible interval is much better than the confidence interval derived from asymptotic theory, thus we will focus on the Bayesian inference.

5.2 Bayesian Inference

5.2.1 Metropolis-Hasting through Asymmetric Laplace Distribution

In Bayesian framework, asymmetric Laplace distribution (Yu and Moyeed 2001, Koenker and Bassett 1978) can be used, whose density function is defined as:

$$f(x|p) = p(1-p)exp(\rho_p(x)),$$

The case when $p = 1/2$ corresponds to a standard Laplace distribution.

For standard linear regression, the error term (i.e. $y_i - X_i\beta$) follows normal distribution. Yu and Moyeed (2001) noted that under the context of quantile regression, it is more convenient to use asymmetric Laplace distribution. Minimizing $\sum_{i=1}^{n} \rho_p(y_i - X_i^T\beta)$ is equivalent to maximize the likelihood function formed by n independent asymmetric Laplace distribution densities. Thus the estimator $\tilde{\beta}$ can be seen as the Maximum Likelihood Estimator, although the indifferentiable nature of $\rho(x)$ prevents people from obtaining a closed-form solution.

Moreover, Sriram et al.(2013) claimed that the use of asymmetric Laplace distribution is satisfactory even if it is not the true underlying distribution.

The posterior of β is:

$$f(\beta|y) \propto L(\beta|y)\pi(\beta)$$

Since the $L(\beta|y)$ has been assumed to be asymmetric Laplace distribution, the remaining problem is the choice of $\pi(\beta)$. Unfortunately, currently there is no conjugate prior has been discovered. Yu and Moyeed (2001) applied noninformative prior (i.e. $\pi(\beta) \propto 1$) and prove that this improper prior leads to proper posterior. Under this setting, the posterior is proportional to:

$$p^n(1-p)^n exp(-\sum_{i=1}^{n} \rho_p(y_i - x_i\beta))$$

The Metropolis-Hasting algorithm could be applied to obtain posterior sample.

5.2.2 Gibbs Sampler through Mixture Representation

Kozumi and Kobayashi (2011) have developed a different way for the Gibbs Sampling of quantile regression, which is based on the location-scale mixture representation of asymmetric Laplace distribution.

Let z be a standard exponential variable and u a standard normal variable. If a random

variable ϵ follows an asymmetric Laplace distribution, then it can be written as:

$$\epsilon = \theta z + \tau\sqrt{z}u,$$
$$\text{where } \theta = \frac{1-2p}{p(1-p)} \text{ and } \tau = \sqrt{\frac{2}{p(1-p)}}$$

Therefore, the response y_i can be written as: $y_i = x_i^T\beta + \theta z_i + \tau\sqrt{z_i}u_i$. If we condition on β and z_i, then y_i follows a normal distribution.

$$f(y|\beta, z) \propto (\prod_{i=1}^{n} z_i^{-\frac{1}{2}}) exp(-\sum_{i=1}^{n} \frac{(y_i - x_i^T\beta - \theta z_i)^2}{2\tau^2 z_i})$$

If we assume the prior of β is $N(\beta_0, B_0)$, then the full conditional of β is:

$$f(\beta|z, y) \propto \pi(\beta) f(y|\beta, z)\cdot$$

Thus, $f(\beta|z, y)$ is also a normal distribution.

$$\beta|z, y \sim N(\beta_p, B_p),$$
$$\text{where } \beta_p = B_p(\sum_{i=1}^{n} \frac{x_i(y_i - \theta z_i)}{\tau^2 z_i} + B_0^{-1}\beta_0) \text{ and } B_p = (\sum_{i=1}^{n} \frac{x_i x_i^T}{\tau^2 z_i} + B_0^{-1})^{-1}$$

As we mentioned before, z_i has an exponential prior with λ=1, hence we have:

$$f(z|\beta, y) \propto \pi(z) f(y|\beta, z)$$
$$\propto z_i^{-\frac{1}{2}} exp\{-\frac{1}{2}(\frac{(y_i - x_i^T\beta)^2 z_i^{-1}}{\tau^2} + \frac{(2\tau^2 + \theta^2)z_i}{\tau^2})\}$$

The full conditional of z_i follows a generalized inverse Gaussian distribution. The density of generalized inverse Gaussian is:

$$\frac{(b/a)^{\nu/2}}{2K_\nu(ab)} x^{\nu-1} exp\{-\frac{1}{2}(a^2 x^{-1} + b^2 x)\},$$

where $K_\nu(x)$ is a modified Bessel function of the third kind. Several software packages have been developed to sample from a generalized inverse Gaussian distribution.

The advantage of this algorithm is that it does not require Metropolis-Hasting procedures so the rejection rate is not an issue to concern about.

5.3 Quantile for Counts

5.3.1 Jittering Technique

To implement quantile regression on count data might cause some problems. The definition of p-quantile for a discrete random variable Y is:

$$q_p = min\{y : P(Y \leq y) \geq p\}$$

Machado and Silva (2005) claimed that the discrete nature of this quantile could lead to invalid asymptotic properties, and the "jittering" technique can effectively solve the problem. Define a new random variable which is:

$$Z = Y + U, U \sim uniform(0, 1)$$

If we denote $P(Y = k) = p_k, k \geq 0$, then the cumulative distribution function of Z is:

$$F(z) = \begin{cases} p_0 z & \text{if } 0 \leq z < 1 \\ \sum_{i=0}^{k-1} p_i + p_k(z-k) & \text{if } k \leq z < z+1 \end{cases}$$

The quantile function of Z is:

$$q(p) = \begin{cases} \frac{p}{p_0} & \text{if } 0 \leq p < p_0 \\ k + \frac{p - \sum_{i=0}^{k-1} p_i}{p_k} & \text{if } \sum_{i=0}^{k-1} p_i \leq p < \sum_{i=0}^{k} p_i \end{cases}$$

However, it is not appropriate to model Z directly by the linear form: $q_z(p|X) = X^T\beta$, because the quantity on the right could be any number in the real line. The p-quantile of response Z should have a lower bound as p. Hence a proper model could be:

$$q_z(p|X) = exp(X^T\beta) + p$$

The quantile is invariant to the monotone transformation, so it is possible to "recover" the linear form. Define:

$$z^* = \begin{cases} log(z - p) & \text{if } p < z \\ log(\delta) & \text{otherwise} \end{cases}$$

δ is a positive number satisfies $0 < \delta < \min\{|z_i - p|, i = 1, ..., n\}$

The new model can be written as:

$$q_{z^*}(p|X) = X^T\beta$$

The β can be estimated by:

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho_p(z_i^* - X_i^T\beta)$$

It is easy to prove that $q_y(p|X) = [q_z(p|X) - 1]$. However, the estimation and interpretation of β are based on z rather than y so it is possible that some β_j is significant for z but not so for y.

5.3.2 Smoothed Check Function

The check function defined in quantile regression is used as a loss function to obtain the M-estimator (Huber 1964). Given the p-quantile, the check function has such form:

$$\rho(x|p) = p * x * I_{[0,+\infty)}(x) - (1-p) * x * I_{(-\infty,0)}(x)$$

The p-quantile regression model obtains the estimate such that:

$$\tilde{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - X_i^T \beta|p)$$

When p=0.5, the check function leads to the estimate of median. As noted by Huber, the derivative of this loss function is called ψ function. The ψ function of the check function in quantile regression is:

$$\psi_\rho(x|p) = \begin{cases} p & \text{if } x \geq 0 \\ p-1 & \text{if } x < 0 \end{cases}$$

The $\rho(x|p)$ is indifferentiable at zero so the value of ψ function at zero has to be determined manually.

It can be proved that the ψ function is proportional to the influence function (Hampel et al. 2011). The shape of the $\psi_\rho(x|p)$ indicates that the extreme values may not affect the estimate too much. On the other hand, the mean function is corresponding to quadratic loss, which yields linear ψ function, so it is not resistant to outliers.

In this paper, we propose a modified check function of the quantile regression whose ψ

function is as follow:

$$\psi_{\rho_2}(x|p,b) = \begin{cases} x+p+bp-1 & \text{if } x \leq -bp \\ p-1 & \text{if } -bp < x < 0 \\ p & \text{if } 0 \leq x < b(1-p) \\ x+p+bp-b & \text{if } x \geq b(1-p) \end{cases}, \text{ where } 0 < p < 1 \text{ and } b \geq 0$$

The $\psi_{\rho_2}(x|p,b)$ has one tuning parameter b which can be seen as the window width. Within the interval $(-bp, b(1-p))$, $\psi_{\rho_2}(x|p,b)$ is exactly the same as $\psi_\rho(x|p)$ so $\psi_{\rho_2}(x|p,b)$ converges to $\psi_\rho(x|p)$ as b goes to infinity. However, the $\psi_{\rho_2}(x|p,b)$ mimics the ψ function of mean outside of that interval. In the whole range, the modified check function is a convex function and common techniques can be used in programming.

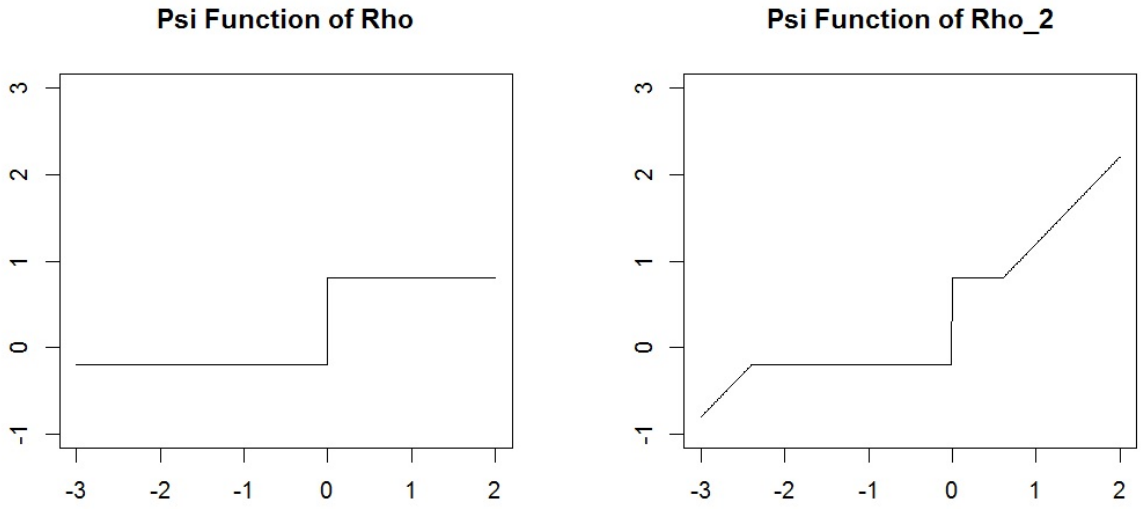

Figure 5.2: Comparison of two check functions with p=0.8 and b=3

Given the ψ function, we can derive the form of $\rho_2(x|p,b)$:

$$\rho_2(x|p,b) = \begin{cases} 0.5x^2 + (p+bp-1)x + b^2p^2/2 & \text{if } x \leq -bp \\ (p-1)x & \text{if } -bp < x < 0 \\ px & \text{if } 0 \leq x < b(1-p) \\ 0.5x^2 + (p+bp-b)x + b^2(1-p)^2/2 & \text{if } x \geq b(1-p) \end{cases}$$

The intercepts are added to ensure that the ρ_2 function to be continuous (as long as $0 < b$).

The above can also be written as:

$$\rho_2(x|p,b) = \begin{cases} \frac{1}{2}(x-bp)^2 + (p-1)x & \text{if } x \leq -bp \\ (p-1)x & \text{if } -bp < x < 0 \\ px & \text{if } 0 \leq x < b(1-p) \\ \frac{1}{2}(x-b(1-p))^2 + px & \text{if } x \geq b(1-p) \end{cases}$$

The form of loss function indicates that: If x is positive, then the loss is increasing as x is larger. If x is negative, then the loss is increasing as -x is larger. Thus, suppose we use an intercept model to estimate the count data, which is nonnegative, then the estimate must be nonnegative as well.

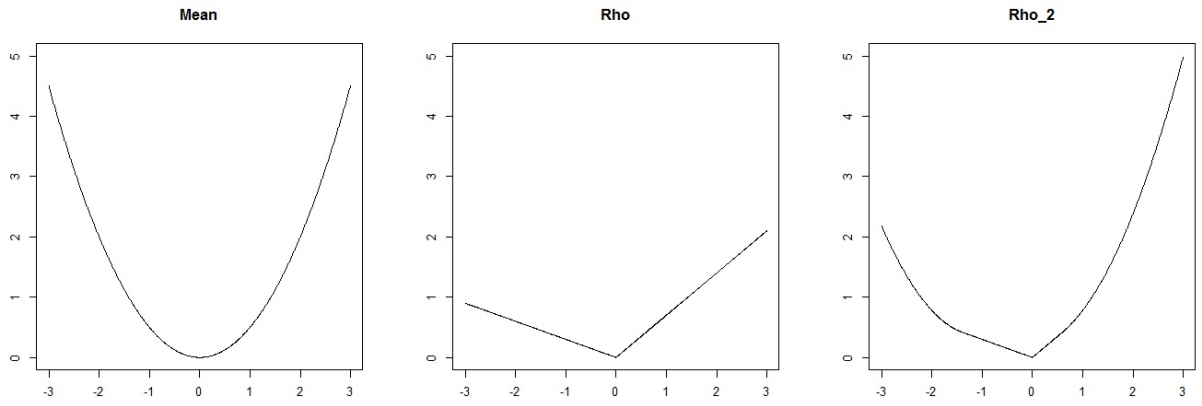

Figure 5.3: Comparison of three loss functions with p=0.7 and b=2

5.3.3 Tuning Parameter: $b \to 0$

A very interesting fact of the $\psi_{\rho_2}(x|p,b)$ function is that when b is close to zero, the loss function turns to be a mixture of quadratic and the p-quantile loss:

$$\psi_{\rho_2}(x|p, b \to 0^+) = \begin{cases} 0.5x^2 + px & \text{if } x \geq 0 \\ 0.5x^2 + (p-1)x & \text{if } x < 0 \end{cases}$$

We will discuss about the equation when $b = 0$ holds. For simplicity, we only consider the intercept model.

Define:

$$\begin{aligned} g(u) = & \, p \int_u^{u+b(1-p)} (y-u) dF_y + (p-1) \int_{u-bp}^{u} (y-u) dF_y \\ & + \int_{u+b(1-p)}^{+\infty} \frac{1}{2}(y-u)^2 + (p+bp-b)(y-u) + \frac{b^2(1-p)^2}{2} dF_y \\ & + \int_{-\infty}^{u+bp} \frac{1}{2}(y-u)^2 + (p+bp-1)(y-u) + \frac{b^2 p^2}{2} dF_y \end{aligned}$$

If we set $b = 0$ and $g'(u) = 0$,

$$g(u) = \int_u^{+\infty} \frac{1}{2}(y-u)^2 + p(y-u)dF_y + \int_{-\infty}^{u} \frac{1}{2}(y-u)^2 + (p-1)(y-u)dF_y$$

$$\Longrightarrow g'(u) = \int_u^{+\infty} (u-y) - pdF_y + \int_{-\infty}^{u} (u-y) - (p-1)dF_y$$

$$= \int_{-\infty}^{+\infty} -ydF_y + (u-p)\int_{-\infty}^{+\infty} dF_y + \int_{-\infty}^{u} dF_y$$

$$= -E(y) + u - p + F_y(u) = 0$$

We will illustrate the use of the above equation by some certain distributions. Assume y follows $U(a,b)$, then the equation within (a,b) becomes:

$$\frac{u-a}{b-a} + u - p - \frac{a+b}{2} = 0$$

Thhe relationship between u and p is **linear**.

Assume y follows $N(\mu, \sigma^2)$, the Taylor expansion of $F(u)$ at μ is:

$$F(u) = F(\mu) + F'(\mu)(u-\mu) + F''(\mu)(u-\mu)^2/2 + R$$

$$(1)F(\mu) = 0.5$$

$$(2)F'(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Longrightarrow F'(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}$$

$$(3)F''(x) = \frac{-(x-\mu)}{\sqrt{2\pi\sigma^2}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Longrightarrow F''(\mu) = 0$$

$$\Longrightarrow F(u) \approx 0.5 + \frac{1}{\sqrt{2\pi\sigma^2}}(u-\mu)$$

Therefore, the equation near $u = \mu$ can be written as:

$$-\mu + u - p + 0.5 + \frac{1}{\sqrt{2\pi\sigma^2}}(u-\mu) = 0$$

The relationship between u and p also has approximately **linear** form.

We have also tested the data set from CEI, which is the number of crashes in 2012 with $E(y) = 0.2371943$. According to our empirical estimate, $F_y(0) = 0.8027819$ and $F_y(1) = 0.9665756$, and actually $F_y(x) = F_y(0)$ for $\forall x \in [0, 1)$. As we talked before, the response y is nonnegative so the estimated value u should also be nonnegative. F_y is not continuous, thus it is possible that we can not find an exact u solves the equation but it is interesting to see how close we can get.

Since $u \geq 0$ and F_y is nondecreasing, the lower bound for $F_y(u) - E(y)$ is $F_y(0) - E(y) = 0.5655876$. The equation prefers that $p - u$ is close to this value. Thus, as long as $p < 0.5655876$, the best u we can get is simply $\hat{u} = 0$. Once p goes beyond that value, the estimate of u will be exactly $\hat{u} = p - 0.5655876$. Therefore, if we draw a plot to indicate the relationship between p and $\hat{u}$, the curve should have two parts: at the beginning $\hat{u}$ stays at 0, and then it increases linearly. The plot below shows the optimization result calculated by software, which is exactly what we expect.

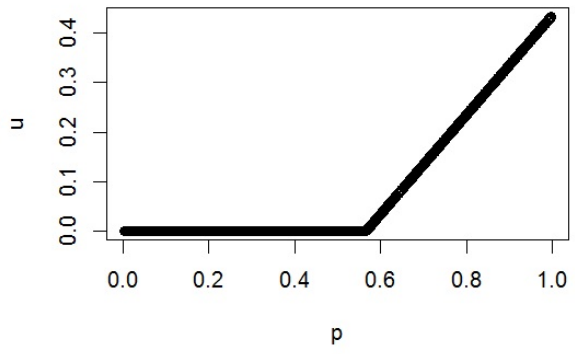

Figure 5.4: The Relationship Between p and u for CEI: Crash in 2012

When b increases, the smoothed check function will yield values closer to the empirical quantile. We will illustrate this by a real simple example. The 100-car data set contains the crashes counts for 108 drivers. The whole data set can be shown as below:

Table 5.1: Crash Data Summary for 100 Car

Number of Crashes	0	1	2	3	4	5
Frequency	73	22	9	2	1	1

The plots below show the relatioship between p and u for different values of b.

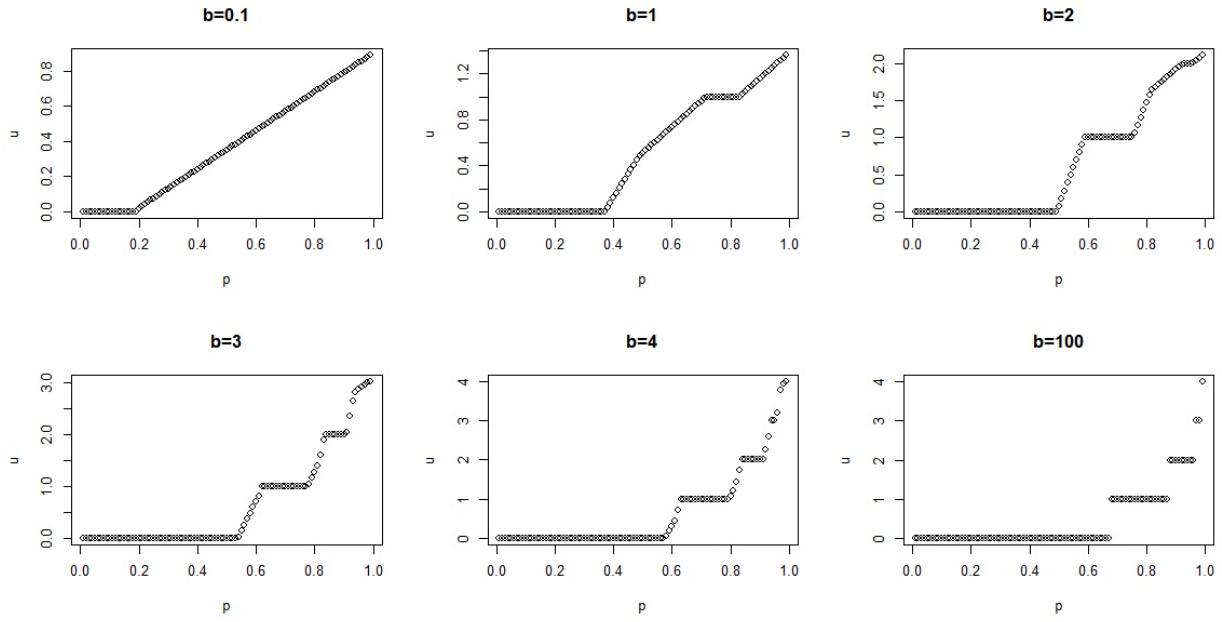

Figure 5.5: The Relationship Between p and u for Different Values of b

5.3.4 Application: 100 Car Data

We also applied the proposed model to the 100 car data and showed that the model estimation based on smoothed check function could be more consistent than the result from ordinary quantile regression.

The quantiles to identify high-risk drivers were 95%, 98% and 99%. By R function "quantile()" we are able to obtain the three empirical quantiles for the number of accidents: 2, 3 and 3.93.

Our covariate for the regression model was the driver's age. The ordinary quantile regression provided following results:

Table 5.2: Model Results: Original Check Function

Parameter	p=0.95	p=0.98	p=0.99
β_0	2.74	5.02	6.67
95% C.I. of β_0	(2.36,5.37)	(3.01,7.08)	(3.6,7.16)
β_1	-0.03	-0.054	-0.083
95% C.I. of β_1	(-0.078,-0.020)	(-0.104,-0.028)	(-0.108,-0.036)

The three models indicated quite different parameter estimates.

However, the $\psi_{\rho_2}(x|p,b)$ was able to generate three quite similar models. We choose b as 4 and the model estimation results are:

Table 5.3: Model Results: Smoothed Check Function

Parameter	p=0.95	p=0.98	p=0.99
β_0	4.80	5.075	5.137
95% C.I. of β_0	(2.37,5.64)	(2.46,5.73)	(3.50,5.90)
β_1	-0.057	-0.056	-0.056
95% C.I. of β_1	(-0.085,-0.012)	(-0.084,-0.007)	(-0.086,-0.024)

5.4 Summary

Identifying the high risk drivers of traffic accidents is crucial for the safety management. Many methods are available. Considering that many factors contribute to crash incidences, most of the methods, such as Poisson regression, offer a way to predict and classify high risk drivers based on the relevant factors. Unfortunately, the current practices in driver safety modeling lack the flexibility and capability to handle heterogeneity and other data issues (Qin et al. (2010)).

In this chapter, we illustrated the use of quantile regression for identifying high risk drivers. The quantile regression is a flexible approach in that it is able to take the different locations of a distribution into account. Thus, by providing estimates at different quantile levels, QR has the capacity to capture heterogeneity in data and present a more well-rounded description of the trends. It is especially useful when a particular location of the data is our major concern

since it provides an optimal estimate based on that location.

We advanced the research of quantile regression in count data by means of the smoothed check function, derived from the idea of M-estimator. This method can be seen as an alternative to the "jittering" method. The smoothed check function has a tuning parameter that controls the final estimation and can be adjusted by the need of users.

To sum up, the quantile regression has great potential in the application of transportation studies, which allows us to determine high risk divers with more specificity and to investigate the possible heterogeneity within the data structure.

Chapter 6

Summary

Statistics plays a central role in current transportation study. Statistical models have been widely used to analyze vehicle crashes, travel time, traffic flow and numerous other subjects. Many popular transportation topics can be statistically interpreted as to analyze the latent structure of the data. In this paper we discussed about the travel time reliability, the individual driver risk classification and high-risk driver identification problem. Several statistical methods have been applied, including Bayesian Gaussian mixture model, hidden Markov model, Poisson Mixture model and Quantile regression.

The multi-state model provides a flexible and efficient framework for modeling travel time reliability, especially under complex traffic conditions. Guo et al. (2012) mentioned that the multi-state model outperforms single-state models in congested or near-congested traffic conditions and the advantage is substantial in high traffic volume condition.

Our objective is to quantitatively evaluate the influence of traffic volume on the mixture of two components. Our work advanced the multi-state models by proposing regressions on the proportions and distribution parameters for underlying traffic states. The Bayesian analysis also provides feasible credible intervals for each parameters without asymptotic assumption.

Previous studies usually model the travel time solely without establishing the relationship between travel time and important transportation statistics such as traffic volume. Our model can also be easily extended to include more covariates in either linear or nonlinear forms.

The modeling results indicated that there is a negatively relationship between the proportion of free-flow state and the traffic volume, which confirms the statement raised by Guo et al. (2012) that for low traffic volume condition, there might only exist one travel time state and single-state models will be sufficient. The estimation for the congested state indicates that the travel time under such condition exhibits substantial variability and positively related with traffic volume, which also verifies the phenomenon found by Guo et al. (2012).

We apply the hidden Markov model to the travel time reliability problem in order to accommodate the dependency structure of observations and to understand how the traffic volume influences on the travel time of vehicles.

Regarding the model specification we consider two possible states of the hidden Markov chain, which are "free-flow" and "congested". The parameters and proportions of the two

states are estimated. Moreover, we apply the well-known logit function in the transition matrix to include the covariate of traffic volume. The modeling result shows that the traffic volume has a positive effect on the proportion of "congested" condition as well as the mean parameters of such condition.

We have compared the model fitting of hidden Markov model with that of ordinary mixture model, and the significant improvement has been shown. We also summarized several issues in this model, such as the technique to determine number of components. To sum up, the hidden Markov model is superior to interpret the data without sacrificing model simplicity.

Developing reliable statistical models for analyzing individual driver risk is very important in transportation studies. In order to deal with the overdispersion issues in crash data, which usually generated from a heterogeneous population, several models have been proposed by the researchers. Unlike the negative binomial model, finite mixture Poisson model is based on the premise that the individual driver risk is dominated by the underlying characteristics, which can be described as "high risk" and "low risk", and may have some relationship with demographic information and historical accident records.

Two levels of uncertainty are quantitatively assessed in the finite mixture Poisson model. The first level of uncertainty is the proportion of a given driver group; the second level of uncertainty is the variation of individual risk within each group. EM algorithm and MCMC method can be used to estimate the model.

Our work provides a comprehensive discussion regarding the limitation of the finite mixture Poisson model. The difficulty of classification can be measured by a simple and intuitive concept: overlap probability. A lower bound of the misclassification rate has been proposed and simulation data are used to justify it. We proved that the overlap probability has a monotonic relationship when the parameter of the second component is changed. We also provided an easy way to find the maximum of overlap probability when the proportion varies. CEI data have been used to illustrate some of the theoretical results.

Identifying the high risk drivers of traffic accidents is crucial for the safety management. Many methods are available. Considering that many factors contribute to crash incidences, most of the methods, such as Poisson regression, offer a way to predict and classify high risk drivers based on the relevant factors. Unfortunately, the current practices in driver safety modeling lack the flexibility and capability to handle heterogeneity and other data issues (Qin et al. (2010)).

We illustrated the use of quantile regression for identifying high risk drivers. The quantile regression is a flexible approach in that it is able to take the different locations of a distribution into account. Thus, by providing estimates at different quantile levels, QR has the capacity to capture heterogeneity in data and present a more well-rounded description of the trends. It is especially useful when a particular location of the data is our major concern since it provides an optimal estimate based on that location.

We advance the research of quantile regression in count data by means of the smoothed check function, derived from the idea of M-estimator. This method can be seen as an alternative to

the "jittering" method. The smoothed check function has a tuning parameter that controls the final estimation and can be adjusted by the need of users.

To sum up, the paper provides a comprehensive overview for dealing with the transportation study by the statistical methods dealing with latent class model, and the prospective problems have been addressed for further research.

Bibliography

[1] Akaho, S. (2008). *Dimension Reduction for Mixtures of Exponential Families*, pages 1–10. Springer.

[2] Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

[3] Albert, P. S. (1991). A two-state markov mixture model for a time series of epileptic seizure counts. *Biometrics*, pages 1371–1381.

[4] Alhamzawi, R., Yu, K., and Benoit, D. F. (2012). Bayesian adaptive lasso quantile regression. *Statistical Modelling*, 12(3):279–297.

[5] Archambeau, C., Lee, J. A., and Verleysen, M. (2003). On convergence problems of the em algorithm for finite gaussian mixtures. In *In Proc. 11th European Symposium on Artificial Neural Networks*, pages 99–106.

[6] Baetschmann, G. and Winkelmann, R. (2013). Modeling zero-inflated count data when exposure varies: With an application to tumor counts. *Biometrical Journal*, 55(5):679–686.

[7] Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997). Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association*, 92(440):1375–1386.

[8] Barrodale, I. and Roberts, F. D. K. (1973). An improved algorithm for discrete l1 linear approximation. *SIAM Journal on Numerical Analysis*, 10(5):839–848.

[9] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563.

[10] Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171.

[11] Baur, D. G. and Lucey, B. M. (2010). Is gold a hedge or a safe haven? an analysis of stocks, bonds and gold. *Financial Review*, 45(2):217–229.

[12] Bearden, W. O., Sharma, S., and Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research*, pages 425–430.

[13] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(7):719–725.

[14] Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the

em algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics Data Analysis*, 41(34):561–575.

[15] Bilmes, J. A. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510):126.

[16] Bliss, C. I. and Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics*, 9(2):176–200.

[17] Bondell, H. D., Reich, B. J., and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika*, 97(4):825–838.

[18] Buchinsky, M. (1994). Changes in the us wage structure 1963-1987: Application of quantile regression. *Econometrica: Journal of the Econometric Society*, pages 405–458.

[19] Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1(8):412–420.

[20] Cameron, A. C. and Trivedi, P. (2013). *Regression analysis of count data*, volume 53. Cambridge University Press.

[21] Celeux, G. and Durand, J.-B. (2008). Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564.

[22] Chang, I. and Kim, S. W. (2012). Modelling for identifying accident-prone spots:

Bayesian approach with a poisson mixture model. *KSCE Journal of Civil Engineering*, 16(3):441–449.

[23] Chen, J. and Cheng, P. (1997). On testing the number of components in finite mixture models with known relevant component distributions. *Canadian Journal of Statistics*, 25(3):389–400.

[24] Chen, L., Baker, S. P., Braver, E. R., and Li, G. (2000). Carrying passengers as a risk factor for crashes fatal to 16- and 17-year-old drivers. *JAMA*, 283(12):1578–1582. 10.1001/jama.283.12.1578.

[25] Chernozhukov, V. (2005). Extremal quantile regression. *Annals of Statistics*, pages 806–839.

[26] Choongrak, K. (1959). A note on box-cox transformation diagnostics. *Technometrics*, 38(2):178.

[27] Chung, H., Walls, T., and Park, Y. (2007). A latent transition model with logistic regression. *Psychometrika*, 72(3):413–435.

[28] Church, K. W. and Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, 1(02):163–190.

[29] Cochrane, D. and Orcutt, G. H. (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association*, 44(245):32–61.

[30] Couvreur, C. (1996). Hidden markov models and their mixtures. *Dept. Math., Universit Catholique de Louvain, Louvain, Belgium*.

[31] Dabrowska, D. M. (1992). Nonparametric quantile regression with censored data. *Sankhy: The Indian Journal of Statistics, Series A*, pages 252–259.

[32] Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in poisson regression models. *Journal of the American Statistical Association*, 84(406):467–472.

[33] Dellaportas, P., Karlis, D., and Xekalaki, E. (1997). Bayesian analysis of finite poisson mixtures.

[34] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

[35] Duijn, M. A. J. v. and Bockenholt, U. (1995). Mixture models for the analysis of repeated count data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 44(4):473–485.

[36] Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, pages 409–428.

[37] E. Hauer, J. N. and Lovell, J. (1988). Estimation of safety at signalized intersections. *Transportation Research Record: Journal of the Transportation Research Board*, pages 48–61.

[38] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, volume 57. CRC press.

[39] Eide, E. and Showalter, M. H. (1998). The effect of school quality on student performance: A quantile regression approach. *Economics Letters*, 58(3):345–350.

[40] Emam, E. and Ai-Deek, H. (2006). Using real-life dual-loop detector data to develop new methodology for estimating freeway travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 1959(-1):140–150.

[41] Farcomeni, A. (2012). Quantile regression for longitudinal data based on latent markov subject-specific parameters. *Statistics and Computing*, 22(1):141–152.

[42] Figueiredo, M. A. T. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(3):381–396.

[43] Findley, L., Smith, C., Hooper, J., Dineen, M., and Suratt, P. M. (2000). Treatment with nasal cpap decreases automobile accidents in patients with sleep apnea. *American journal of respiratory and critical care medicine*, 161(3):857–859.

[44] Firpo, S., Fortin, N. M., and Lemieux, T. (2009). Unconditional quantile regressions. *Econometrica*, 77(3):953–973.

[45] Formisano, R., Bivona, U., Brunelli, S., Giustini, M., Longo, E., and Taggi, F. (2005). A preliminary investigation of road traffic accident rate after severe brain injury. *Brain Injury*, 19(3):159–163.

[46] Fowlkes, E. B. (1979). Some methods for studying the mixture of two normal (lognormal) distributions. *Journal of the American Statistical Association*, 74(367):561–575.

[47] Frhwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*.

[48] Gardner, W., Mulvey, E. P., and Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed poisson, and negative binomial models. *Psychological bulletin*, 118(3):392.

[49] Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154.

[50] Ghosh, S., Gelfand, A. E., Zhu, K., and Clark, J. S. (2012). The kzig: Flexible modeling for zeroinflated counts. *Biometrics*, 68(3):878–885.

[51] Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732.

[52] Grn, B. and Leisch, F. (2004). Bootstrapping finite mixture models.

[53] Guo, F., Li, Q., and Rakha, H. (2012). Multistate travel time reliability models with skewed component distributions. *Transportation Research Record: Journal of the Transportation Research Board*, 2315(-1):47–53.

[54] Guo, F., Rakha, H., and Park, S. (2010). Multistate model for travel time reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 2188(-1):46–54.

[55] Guo, J. and Trivedi, P. (2002). Flexible parametric models for long-tailed patent count distributions. *Oxford Bulletin of Economics and Statistics*, 64:63–82.

[56] Hakamies-Blomqvist, L., Raitanen, T., and ONeill, D. (2002). Driver ageing does not cause higher accident rates per km. *Transportation Research Part F: Traffic Psychology and Behaviour*, 5(4):271–274.

[57] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384.

[58] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*, volume 114. Wiley. com.

[59] Harrington, S. E. and Doerpinghaus, H. I. (1993). The economics and politics of automobile insurance rate classification. *The Journal of Risk and Insurance*, 60(1):59–84.

[60] Harter, S. P. (1975). A probabilistic approach to automatic keyword indexing. part i. on the distribution of specialty words in a technical literature. *Journal of the American Society for Information Science*, 26(4):197–206.

[61] Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *Journal of the American Statistical Association*, 64(328):1459–1471.

[62] Hauer, E. (2001). Overdispersion in modelling accidents on road sections and in empirical bayes estimation. *Accident Analysis Prevention*, 33(6):799–808.

[63] Heydecker, B. and Wu, J. (2001). Identification of sites for road accident remedial work by bayesian statistical methods: an example of uncertain inference. *Advances in Engineering Software*, 32(10):859–869.

[64] Horowitz, J. L. and Lee, S. (2005). Nonparametric estimation of an additive quantile regression model. *Journal of the American Statistical Association*, 100(472):1238–1249.

[65] Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

[66] Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

[67] Jaccard, J. and Turrisi, R. (2003). *Interaction effects in multiple regression*. Sage.

[68] Jansakul, N. and Hinde, J. (2002). Score tests for zero-inflated poisson models. *Computational statistics data analysis*, 40(1):75–96.

[69] Jean-Luc, G. and Chin-Hui, L. (1991). Bayesian learning of gaussian mixture densities for hidden markov models. 112457 272-277.

[70] Jinyong, H. (1995). Bootstrapping quantile regression estimators. *Econometric Theory*, 11(1).

[71] John, F. G. and Michael, P. K. (1997). Mixture of normals probit models. Technical report, Federal Reserve Bank of Minneapolis. Staff Report.

[72] Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

[73] Jovanis, P. P. and Chang, H.-L. (1986). *Modeling the Relationship of Accidents to Miles Traveled*.

[74] Karlis, D. and Xekalaki, E. (2003). Choosing initial values for the em algorithm for finite mixtures. *Computational Statistics and Data Analysis*, 41(34):577–590.

[75] Keating, D. P. (2007). Understanding adolescent development: Implications for driving safety. *Journal of Safety Research*, 38(2):147–157.

[76] Koenker, R. (1994). *Confidence intervals for regression quantiles*, pages 349–359. Springer.

[77] Koenker, R. (2006). *Quantile Regresssion*. John Wiley Sons, Ltd.

[78] Koenker, R. and Bassett, Gilbert, J. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

[79] Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283.

[80] Kottas, A. and Krnjaji, M. (2009). Bayesian semiparametric modelling in quantile regression. *Scandinavian Journal of Statistics*, 36(2):297–319.

[81] Koutsourelis, A. and Yu, K. (2012). *Bayesian Extreme Quantile Regression for Hidden Markov Models*. Brunel University.

[82] Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of statistical computation and simulation*, 81(11):1565–1578.

[83] Krogh, A., Brown, M., Mian, I. S., Sjlander, K., and Haussler, D. (1994). Hidden markov models in computational biology: Applications to protein modeling. *Journal of molecular biology*, 235(5):1501–1531.

[84] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

[85] Laapotti, S., Keskinen, E., and Rajalin, S. (2003). Comparison of young male and female drivers' attitude and self-reported traffic behaviour in finland in 1978 and 2001. *Journal of Safety Research*, 34(5):579–587.

[86] Lajunen, T. and Parker, D. (2001). Are aggressive people aggressive drivers? a study of the relationship between self-reported general aggressiveness, driver anger and aggressive driving. *Accident Analysis Prevention*, 33(2):243–255.

[87] Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

[88] Lancaster, T. and Jae Jun, S. (2010). Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25(2):287–307.

[89] Langford, J., Methorst, R., and Hakamies-Blomqvist, L. (2006). Older drivers do not have a high crash riska replication of low mileage bias. *Accident Analysis Prevention*, 38(3):574–578.

[90] Lenk, P. and DeSarbo, W. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65(1):93–119.

[91] Li, G., Baker, S. P., Langlois, J. A., and Kelen, G. D. (1998). Are female drivers safer? an application of the decomposition method. *Epidemiology*, 9(4):379–384.

[92] Li, J. and Gray, R. M. (2000). *Image segmentation and compression using hidden Markov models*. Springer.

[93] Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3):533–556.

[94] Li, W. and Carriquiry, A. (2005). *The Effect of Four-Lane to Three-Lane Conversion on the Number of Crashes and Crash Rates in Iowa Roads*. Department of Statistics, Iowa State University.

[95] Lim, H. K., Li, W. K., and Yu, P. L. H. (2013). Zero-inflated poisson regression mixture model. *Computational Statistics Data Analysis*, (0).

[96] Lo, Y., Mendell, N. R., and Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3):767–778.

[97] Lord, D. and Bonneson, J. A. (2005). Calibration of predictive models for estimating safety of ramp design configurations. *Transportation Research Record: Journal of the Transportation Research Board*, 1908(1):88–95.

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

[99] Lord, D., Washington, S., and Ivan, J. N. (2007). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis Prevention*, 39(1):53–57.

[100] Lord, D., Washington, S. P., and Ivan, J. N. (2005). Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. *Accident Analysis Prevention*, 37(1):35–46.

[101] MacEachern, S. N. and Mller, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.

[102] Machado, J. A. F. and Silva, J. M. C. S. (2005). Quantiles for counts. *Journal of the American Statistical Association*, 100(472):1226–1237.

[103] MacKay Altman, R. (2004). Assessing the goodnessoffit of hidden markov models. *Biometrics*, 60(2):444–450.

[104] Marin, J.-M., Mengersen, K. L., and Robert, C. (2005). *Bayesian modelling and inference on mixtures of distributions*. Elsevier. For more information about this book please refer to the publisher's website (see link) or contact the author . Author contact details : k.mengersen@qut.edu.au.

[105] Massie, D. L., Campbell, K. L., and Williams, A. F. (1995). Traffic accident involvement rates by driver age and gender. *Accident Analysis and Prevention*, 27(1):73–87.

[106] Maze, T., Agarwai, M., and Burchett, G. (2006). Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. *Transportation Research Record: Journal of the Transportation Research Board*, 1948(-1):170–176.

[107] McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied Statistics*, pages 318–324.

[108] Meeker, W. Q. and Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician*, 49(1):48–53.

[109] Mohri, M. and Roark, B. (2005). Structural zeros versus sampling zeros. Report, Technical Report CSE-05-003, Computer Science Electrical Engineering, Oregon Health Science University.

[110] Monrrez-Espino, J., Hasselberg, M., and Laflamme, L. (2006). First year as a licensed car driver: Gender differences in crash experience. *Safety Science*, 44(2):75–85.

[111] Neyman, J. and Pearson, E. S. (1933). *On the problem of the most efficient tests of statistical hypotheses*. Springer.

[112] Ng, K.-s., Hung, W.-t., and Wong, W.-g. (2002). An algorithm for assessing the risk of traffic accident. *Journal of safety research*, 33(3):387–410.

[113] Oja, E. (2002). Unsupervised learning in neural computation. *Theoretical Computer Science*, 287(1):187–207.

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

[115] Park, S., Rakha, H., and Guo, F. (2010). Multi-state travel time reliability model: Model calibration issues. In *89th Transportation Research Board Annual Meeting. Transportation Research Board*.

[116] Park, S., Rakha, H., and Guo, F. (2011). Multi-state travel time reliability model: Impact of incidents on travel time reliability. In *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, pages 2106–2111. IEEE.

[117] Poch, M. and Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering*, 122(2):105–113.

[118] Portnoy, S. and Jureckova, J. (1999). On extreme regression quantiles. *Extremes*, 2(3):227–243.

[119] Qin, X., Ng, M., and Reyes, P. E. (2010). Identifying crash-prone locations with quantile regression. *Accident Analysis Prevention*, 42(6):1531–1537.

[120] Qin, X. and Reyes, P. E. (2011). Conditional quantile analysis for crash count data. *Journal of Transportation Engineering*, 137(9):601–607.

[121] Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

[122] Raikov, D. (1937). On the decomposition of poisson laws. *Comptes Rendus de lAcademie Science de lURSS*, 14:9–11.

[123] Rau, A., Celeux, G., Martin-Magniette, M.-L., and Maugis-Rabusseau, C. (2011). Clustering high-throughput sequencing data with poisson mixture models. Report.

[124] Reeder, A. I., Alsop, J. C., Begg, D. J., Nada-Raja, S., and McLaren, R. L. (1998). A longitudinal investigation of psychological and social predictors of traffic convictions among young new zealand drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 1(1):25–45.

[125] Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5):465–471.

[126] Rydn, T. (2008). Em versus markov chain monte carlo for estimation of hidden markov models: A computational perspective. *Bayesian Analysis*, 3(4):659–688.

[127] Schaeck, K. (2008). Bank liability structure, fdic loss, and time to failure: A quantile regression approach. *Journal of Financial Services Research*, 33(3):163–179.

[128] Schuh, A., Woodall, W. H., and Camelio, J. A. (2013). The effect of aggregating data when monitoring a poisson process. *Journal of Quality Technology*, 45(3):260–272.

[129] Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

[130] Scott, S. L. (2002). Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 97(457):337–351.

[131] Sloan, F. A., Reilly, B. A., and Schenzler, C. (1995). Effects of tort liability and insurance on heavy drinking and drinking and driving. *JL Econ.*, 38:49.

[132] Smyth, P. (1994). Hidden markov models for fault detection in dynamic systems. *Pattern recognition*, 27(1):149–164.

[133] Sriram, K., Ramamoorthi, R., and Ghosh, P. (2013). Posterior consistency of bayesian quantile regression based on the misspecified asymmetric laplace density. *Bayesian Analysis*, 8(2):1–16.

[134] Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components- an alternative to reversible jump methods. *The Annals of Statistics*, 28(1):40–74.

[135] Taddy, M. A. and Kottas, A. (2010). A bayesian nonparametric approach to inference for quantile regression. *Journal of Business Economic Statistics*, 28(3).

[136] Tan, P.-N. (2007). *Introduction to data mining*. Pearson Education India.

[137] Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research*, 178(1):154–167.

[138] Teicher, H. (1960). On the mixture of distributions. *The Annals of Mathematical Statistics*, 31(1):55–73.

[139] Tu, H., Van Lint, J., and van Zuylen, H. J. (2008). Travel time reliability model on freeways. In *Transportation Research Board 87th Annual Meeting*.

[140] Turner, C. and McClure, R. (2003). Age and gender differences in risk-taking behaviour as an explanation for high incidence of motor vehicle crashes as a driver in young males. *Injury control and safety promotion*, 10(3):123–130.

[141] Venezian, E. (1981). Good and bad drivers-a markov model of accident proneness. *Proceedings of the Casualty Actuarial Society Casualty Actuarial Society*, LXVIII:65–85.

[142] Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

[143] Viallefont, V., Richardson, S., and Green, P. J. (2002). Bayesian analysis of poisson mixtures. *Journal of Nonparametric Statistics*, 14(1-2):181–202.

[144] Visser, I., Raijmakers, M. E., and Molenaar, P. (2000). Confidence intervals for hidden markov model parameters. *British journal of mathematical and statistical psychology*, 53(2):317–327.

[145] Visser, I. and Speekenbrink, M. (2010). depmixS4: An R package for hidden markov models. *Journal of Statistical Software*, 36(7):1–21.

[146] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269.

[147] Vounatsou, P., Smith, T., and Smith, A. F. M. (1998). Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(4):575–587.

[148] Wallace, C. S. and Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, 11(2):185–194.

[149] Walters, M. A. (1981). Risk classification standards. *Proceedings of the Casualty Actuarial Society*, 68:1–18.

[150] Wang, C.-P., Hendricks Brown, C., and Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventive intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association*, 100(471):1054–1076.

[151] Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics*, 52(2):381–400.

[152] Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodnessoffit of parametric models to multivariate abundance data. *Environmetrics*, 16(3):275–289.

[153] Weber, D. C. (1970). A stochastic approach to automobile compensation. *Proceedings of the Casualty Actuarial Society*, 57:27–63.

[154] Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gaussnewton method. *Biometrika*, 61(3):439–447.

[155] Wedel, M., DeSarbo, W. S., Bult, J. R., and Ramaswamy, V. (1993). A latent class poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, 8(4):397–411.

[156] Whittaker, J., Whitehead, C., and Somers, M. (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(5):863–878.

[157] Wiener, N. (1930). The auto-correlation function. *Acta. Math*, 55:273.

[158] Willett, A. H. (1901). *The economic theory of risk and insurance*. The Columbia university press.

[159] Winkelmann, R. (2006). Reforming health care: Evidence from quantile regressions for counts. *Journal of Health Economics*, 25(1):131–145.

[160] Wold, H. (1938). A study in the analysis of stationary time series.

[161] Wolfe, J. H. (1971). A monte carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. Report, DTIC Document.

[162] Wu, C. F. J. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1):95–103.

[163] Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19(2):801.

[164] Yang, M.-S. and Lai, C.-Y. (2005). Mixture poisson regression models for heterogeneous count data based on latent and fuzzy class analysis. *Soft Computing*, 9(7):519–524.

[165] Yang, R. and Berger, J. O. (1996). *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University.

[166] Yu, K. (2002). Quantile regression using rjmcmc algorithm. *Computational statistics data analysis*, 40(2):303–315.

[167] Yu, K., Chen, C. W., Reed, C., and Dunson, D. B. (2013). Bayesian variable selection in quantile regression. *Statistics and Its Interface*, 6:261274.

[168] Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American statistical Association*, 93(441):228–237.

[169] Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):331–350.

[170] Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics Probability Letters*, 54(4):437–447.

[171] Yue, Y. R. and Rue, H. (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics Data Analysis*, 55(1):84–96.

[172] Yuting, Q., Paisley, J. W., and Carin, L. (2007). Music analysis using hidden markov mixture models. *Signal Processing, IEEE Transactions on*, 55(11):5209–5224.

[173] Zhang, Z., Chan, K., Wu, Y., and Chen, C. (2004). Learning a multivariate gaussian mixture model with the reversible jump mcmc algorithm. *Statistics and Computing*, 14(4):343–355.

[174] Zhao, S. and Richard Jr, B. (2007). *Planning Chinese characters: reaction, evolution or revolution?*, volume 9. Springer.

[175] Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov models for time series: an introduction using R*. CRC Press.