Data Analytics of Human Behaviors with Applications

by

Matthew-Remy Aguirre

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Jionghua (Judy) Jin, Chair
Assistant Professor Raed Al Kontar
Associate Professor Eunshin Byon
Assistant Professor Yang Chen

Matthew-Remy Aguirre
mattagui@umich.edu
ORCID iD: 0000-0003-3903-4067

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to everyone who has contributed to the completion of this dissertation.

I am deeply grateful for having Dr. Jin as my advisor throughout my PhD, as she has always been willing to share her guidance and insights with me throughout this journey. Her expertise and mentorship have not only shaped the direction of my research, but also helped me grow as an independent researcher. I will always remember (and be thankful for) the amount time we would spend in one-on-one meetings discussing the various projects and methodologies I was currently working on.

I am also grateful to the members of my dissertation committee, Dr. Al Kontar, Dr. Byon, and Dr. Chen. Their thoughtful feedback and constructive suggestions have significantly improved the quality of this work, and helped guide me to other research that relates to my work. I'd like to specifically thank Dr. Chen for encouraging me to take part in her research paper study group, as this gave me a nice opportunity to see the perspectives of PhD students belonging the Statistics department. That environment really helped me grow and network as a researcher.

I also have a sincere appreciation for all my colleagues, fellow students, and fellow workmates who have provided stimulating discussions, encouragement, and a collaborative atmosphere that made the research process enjoyable. I'd like to specifically thank my labmates: Dr. Wenbo Sun always being a friendly face to discuss proofs with, and Jiacheng Liu for always being willing to listen in to my presentations and read over my writing.

I extend my thanks to my family for their boundless love, encouragement, and belief in me. Their unwavering support has been my anchor during the highs and lows of this academic journey. Without my parents' encouragement, I would never have thought to have reached this part of my academic career.

Last but not least, I am grateful to all the teachers, mentors, friends, and everyone who has touched my life in some way, contributing to my personal and academic growth. Specifically, I'd like to thank my significant other, Claudia, for staying with me these past seven years, providing emotional support, and impacting my career path decisions.

In closing, this dissertation stands as a testament to the collective efforts of numerous individuals, and I am honored to acknowledge their contributions.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

This dissertation presents a comprehensive analysis of human behaviors with applications in detecting distractive driving behaviors, automatic labeling of common driving behaviors, and the prediction of consumer behavior related to purchasing airline tickets. The findings aim to foster our understanding of the complexities that drive human behaviors. Emphasis is placed on the development of practical methods that generate a positive societal impact with regards to driving safety and business management decisions.

The insights of human behaviors were explored through three major works: (1) automatic clustering of diverse kinematic driving data patterns to enhance the understanding of drivers' normal maneuvering behaviors; (2) detecting distracted driving behaviors to enhance human driving safety through the fusion of kinematic data and driving models; and (3) modeling and predicting consumers' behaviors to assist business management decisions in pricing airline tickets. The major methodological contributions are summarized as follows.

The first work introduces a new multi-model fusion and detection algorithm through effectively synthesizing three different state-space models, each of which represents a driver's typical kinematic motion pattern (i.e. constant velocity, constant acceleration, and constant turn rate and acceleration). The kinematic motion trajectory can be dynamically updated and predicted by using the Kalman filtering method. A normalized likelihood is further proposed for model fusion by weighing the contribution of each kinematic motion model. The results are then used to build two monitoring control charts. Specifically, both an Exponentially Weighted Moving Average (EWMA) control chart and a Cumulative Sum (CUSUM) control chart are developed to automatically detect and assess distracted driving behavior. This work is attractive in that it can be used to generate real-time warning signals during driving to avoid risky distracted driving behaviors.

The second major work aims to identify typical drivers' normal maneuvering behaviors from a collected naturalistic driving dataset. The automatic clustering of time series kinematic driving data is done by improving the existing Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM) inference method. In literature, utilizing a Hierarchical Dirichlet Process as a prior can lead to consistency issues, which can harm the estimation of duration probabilities, transition probabilities, and the total number of maneuver states. Although many solutions have been proposed to counter the consistency issues, none have been from the perspective of HDP-HSMMs. For this

purpose, this research presents an enhanced robust algorithm called rHDP-HSMM, which adds a state-merging procedure to each iteration of the inference procedure to avoid redundant states. As a result, the proposed rHDP-HSMM leads to more consistent states, faster convergence, and better estimations of the emission and duration parameters than the existing HDP-HSMM method. Therefore, the proposed rHDP-HSMM model can improve the estimation accuracy of the true number of states of a given time series dataset. This enhances the understanding and inference of drivers' maneuvering behaviors. Furthermore, the generated clustering results can be served as a surrogated model to generate realistic simulation data for testing autonomous vehicles.

The third major work combines concepts from economics and constrained Gaussian Process (CGP) regression to probabilistically describe consumer purchasing behaviors. Fully modeling customer buy-down and sell-up behaviors for all different fare groups is often difficult due to the sparse or lack of customers' purchasing data in some fare groups. The existing system-wide aggregation/averaging methods can be used to address this sparsity issue, however, the resulting customer behavior estimates become uncharacteristic of the original markets of interest. This work addresses this issue by applying the multitask learning strategy to a CGP model, in which all fare groups' data are shared to estimate individual fare group demand simultaneously. The proposed CGP model enables an accurate estimation of customers' purchasing behaviors in fare groups that have limited historical purchasing data. The estimates can be further used to infer customer buy-down and sell-up probabilities. Specifically, a general CGP model is developed by adding appropriate airline demand constraints that represent the relationships both within and between fare groups. The proposed CGP model facilitates an inter-group cross-learning capability for predicting customer purchasing behaviors. In addition, it is also proposed to adjust a particular set of airline demand constraints such that the economics concept of price elasticity can be taken into account. This leads to an improved estimation of customer demand. To further consider the buy-down and sell-up behaviors, the estimated demand is decomposed into three different states (cheap, expensive, no longer sold) to consider which fare groups were available at the time of purchase. This decomposition allows for historical observations to represent different types of buying behaviors. A case study is presented to demonstrate how demand estimates of the CGP can be converted into both buy-down and sell-up probabilities that are dependent on selling time. As such, this work offers both modeling contributions and a better understanding of customer purchasing behaviors in airline applications.

Collectively, the contributions from all three works presented in this dissertation have advanced data analytics methods in modeling and inferring human behaviors across the applications of transportation and business management field. The findings also enrich the understanding of human behaviors to foster the development of practical solutions in the corresponding application fields.

CHAPTER 1

Introduction

1.1 Motivation

In recent years, the advancements in data analytical methods have opened up new possibilities for understanding human behaviors in various application domains. This dissertation explores the development of algorithms that delve into distinct but interconnected aspects of human behavior analysis methods. By utilizing advanced statistical modeling and machine learning techniques, this dissertation has advanced data analytics methods for understanding and predicting human distracted driving behavior, naturalistic human driving behaviors, and humanistic buying behavior.

1.1.1 Motivation for Modeling Distracted Driving Behavior

The first major work focuses on the detection of human distracted driving behavior. With the increasing prevalence of mobile devices and other distractions, there is a growing concern for driving safety. There is much motivation for developing algorithms that enable the detection of inattentive driving behaviors. Driver distraction occurs when a driver diverts their attention away from the task of driving to focus on a secondary task. For example, texting is a common secondary task performed while driving that causes driver distraction. Much research has been documented on the topic of driver distraction and can usually be categorized into one of the two different approaches.

The more common approaches focus on monitoring the drivers directly, as drivers themselves often display certain patterns when becoming distracted. For example, these patterns can be the driver reaching for their phone or even displaying an abnormal blink rate. Research based on the this approach is usually performed via techniques involving face and facial landmark detection, biomechanical detection and recognition, and measuring cognitive load. However, there are many practicalities to consider before these techniques can be implemented. For example, additional sensors are often required to be either installed in the vehicle or equipped to drivers directly.

Not only does incur additional costs to the drivers of interest, but issues privacy may result from implementation. Hence in practice, implementation of above type of research is still very limited.

Chapter 2 focuses on analyzing the vehicle's real-time driving patterns. The idea is that any secondary task distracting the driver will directly manifest in their vehicle maneuver control. Hence this approach will focus on using already available kinematic motion signals to predict how situations will evolve and measure a driver's deviation from the predicted behavior. The sensors used in this framework are already embedded in a vehicle measuring information like speed, yaw rate, and acceleration.

Many well-defined motion models compliment this approach. They are termed *motion models* because these models track and predict signals describing the motion of an object. Specifically, Chapter 2 will utilize kinematic motion models make use of the well-understood physics relationships between the signals of interest. For example, the assumptions of constant velocity (CV), constant acceleration (CA), or even constant turn rate and acceleration (CTRA) can be be used to predict the motion of the vehicle. These use the state-space modeling approach in conjunction with kinematic equations to capture dynamic vehicle motion trajectories.

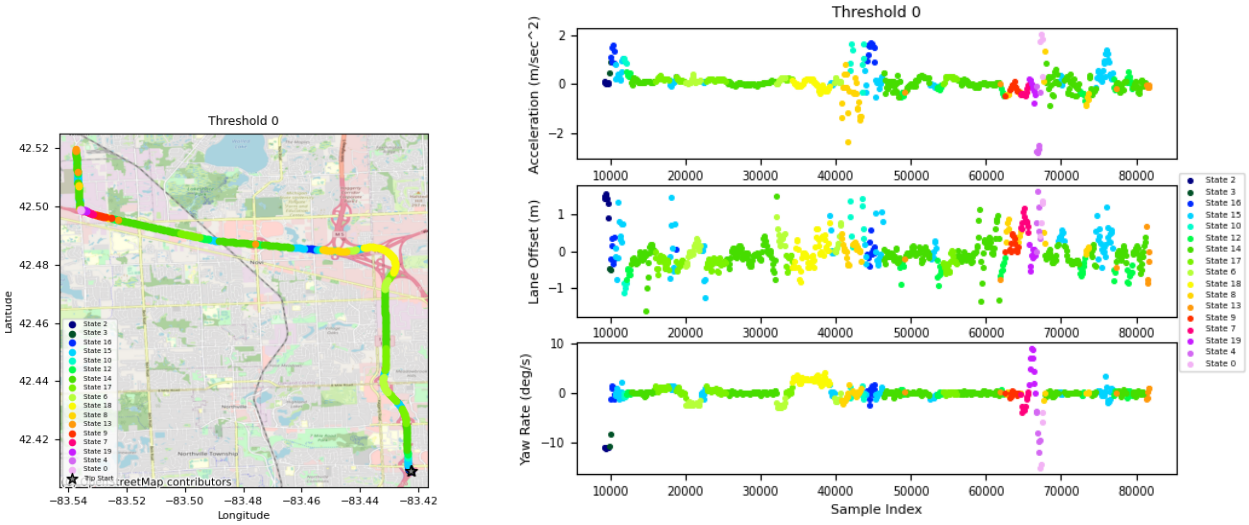
However, not much literature exists relating the use of kinematic motion models to detect driver distraction. One paper that comes close is [57], where four motion models with both the extended Kalman filter and a particle filter are used to detect lane level irregular driving behaviors via fuzzy logic to output risk types. However their use of GPS signal data presented some challenges in the quality of measurement data. Even so, [56] notes that most of the irregular driving behavior based on vehicle-based research is still preliminary and in early stages of development as most literature does not attempt to quantify performance in terms of correct detection rate. As such, in the field of detecting driver distraction, most researchers tend to use more visual based methods described by the first approach.

1.1.2 Motivation for Modeling Naturalistic Human Driving Behaviors

The analysis of vehicle driving styles is prominent to the field of intelligent transportation, vehicle calibration, and autonomous testing [69, 46]. The term *driving style* can be referred as a set of dynamic activities or steps that a driver uses when driving. In other words, vehicle driving styles describe how humans naturalistically perform various driving maneuvers. Hence, this type of research impacts eco-driving, road safety, and intelligent vehicles [7, 49, 35]. Chapter 3 seeks to automatically describe these naturalistic human driving behaviors through the analysis of kinematic signals.

To describe and model these driving styles, one popular approach is the use of a Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM) [62]. This model is powerful in that

it considers the sequential nature of driving kinematic signals, and estimates data segmentation, behavior state duration, and state transition probabilities. The HDP-HSMM provides semantical way for analyzing driver behaviors, and is thus popularly used for describing driving styles. Figure 1.1a shows an exemplar set of sequential kinematic signals belonging to the trip observed in Figure 1.1b. The signals are color-coded to reference a state segmentation determined by a HDP-HSMM.



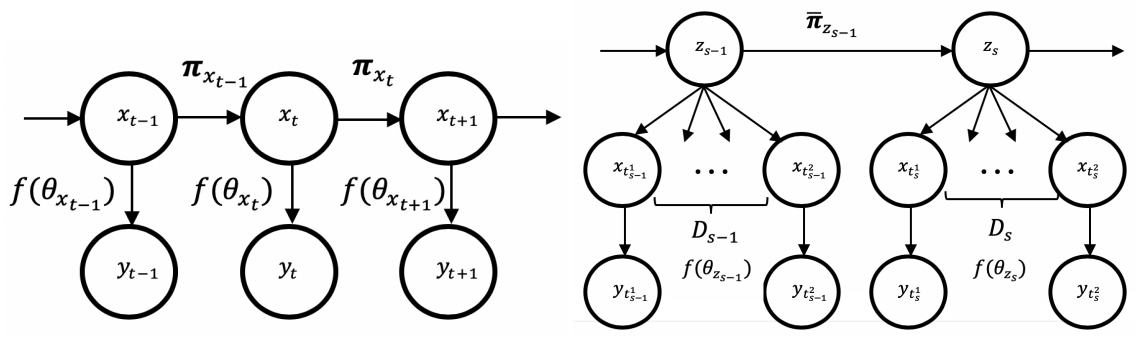
(a) **(b)**

Figure 1.1: *An example trip from Chapter 3 and the kinematic signals belonging to it. Learned states from an HDP-HSMM are color coded as labels.*

The HDP-HSMM is able to describe naturalistic human driving behavior in great detail due to the structure of the model. The HDP-HSMM utilizes a three-layer structure (shown in Figure 1.2a) which demonstrates the effects of a Semi-Markovian approach to model the state transitions $\bar{\pi}_{z_s}$. This structure allows users to model the duration D_s of each behavior via any discrete distribution, while allowing the segmentation of each behavior to be directly represented by the hidden states z_s. Furthermore, the HDP-HSMM is able to infer the number of behaviors represented by data by use of the Hierarchical Dirichlet Process (HDP). The HDP acts a prior on the transition probabilities to allow which allows for information to be shared between the different state transition probabilities. Utilizing this process as a prior to the transition probabilities translates to the number of states also being determined via a Dirichlet Process type of procedure.

While the HDP-HSMM is powerful, literature outside of the field of transportation details how the model's use of an HDP prior can lead to redundant and inconsistent state estimations. This detail is important as it needs to be considered by researchers attempting to utilize the HDP-HSMM to describe driving styles. For example, Figure 1.1b clearly has redundant states as seen by the green

shaded states. The redundant states can make analysis of HDP-HSMM outputs across multiple datasets difficult for researchers hoping to model naturalistic human driving behaviors.



(a) *HMM* **(b)** *HSMM*

Figure 1.2: *A comparison between the structure of a Hidden Markov Model (HMM) and a Hidden Semi-Markov Model (HSMM). The variables and their descriptions are as follows: x_t (hidden state at time t), y_t (observed data at time t), π_x (transition probabilities of state x), $f(\theta_x)$ (probability distribution of state x), z_s (state of segment s), D_s (state duration of segment s).*

A few works exist that focus on solving this issue for the more simpler Hidden Markov Model (displayed in Figure 1.2a). [15] discussed HMM's utilizing a Dirichlet prior, and the assumptions on the prior required for the consistency. [11] developed the *sticky* HDP-HMM (sHDP-HMM) to consider the issue of redundant states. This model adds a bias to the prior on the rows of the transition matrix which emphasizes self-transitions. This results in an increased state duration for each learnt state, which allows the sHDP-HMM to avoid redundant states with short state duration. However, this strategy cannot be applied to HDP-HSMM as the modeling structure of HMM's is inherently different from HSMM's. Outside of HMM and HSMM modeling, [18] focused on the Dirichlet Process Mixture model, and presented the Merge-Truncate-Merge algorithm, which guaranteed a consistent estimate to the number of mixture components. This post-processing procedure takes advantage of the fact that the posterior sample tends to produce a large number of atoms with small weights, and probabilistically merges atoms together.

Given these approaches, Chapter 3 attempts to address the HDP's inconsistency problem by taking inspiration from both the sticky HDP-HMM and the Merge-Truncate-Merge algorithm.

1.1.3 Motivation for Modeling Humanistic Buying Behaviors

In the field of revenue management, understanding consumer behavior is important for capacity control decisions. Capacity control decisions are any decision impacting the allocation of capacity of a resource to various consumers. For airline revenue management, airlines consider how much capacity should be allocated to various fare groups before deciding the price of each seat on a

plane. This decision impacts the various shopping options customers see when purchasing, which in turn impacts consumer buying behavior. Customers will respond differently depending on the options presented before them. The more relevant a product appears to a customer, the higher the probability of a purchase being made. For example, if an option is within their preferred price range, a customer may make a purchase. If a product is slightly above their preferred price range, there is still a chance that a customer may make a purchase. As such, airlines must consider how their decisions regarding which fare groups to show the customers affect customer buying behavior.

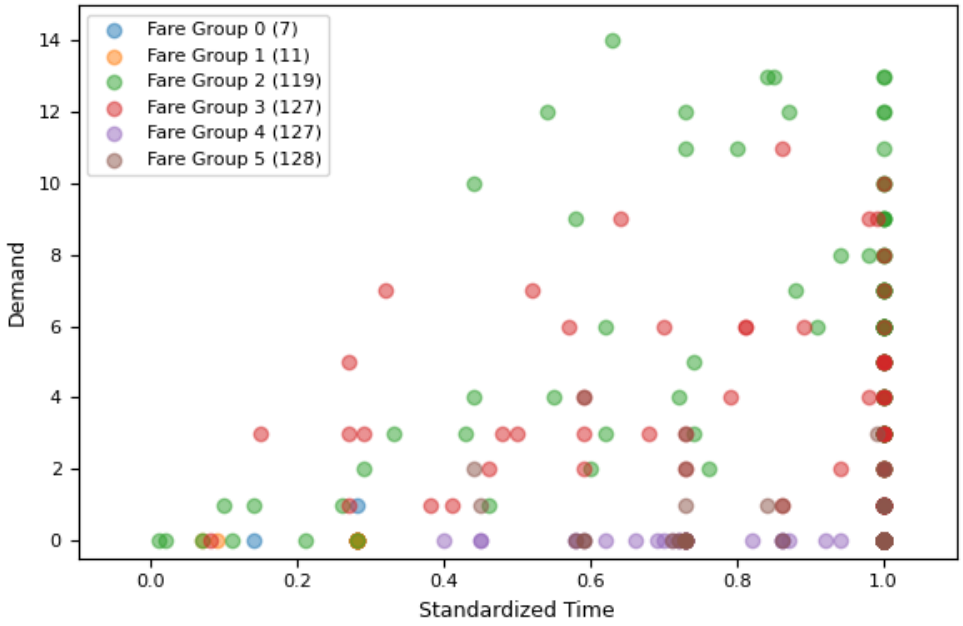


Figure 1.3: *Example of airline booking data from different fare groups. Each group is represented by a different color. The number of samples collected for each fare group is shown in the legend in parenthesis.*

The term *customer buying behavior* implies modeling demand dependencies between different fare groups. One method for modeling demand dependencies between fare groups is the use of sell-up rate probabilities. These probabilities are meant to capture the probability of customers 'selling-up' between the fares. For example, [21] utilized fare ratios and observed demand to estimate sell-up rates. [14] provides an alternative sell-up rate calculation using a multinomial logit model to consider cross-flight effects. Sell-up rates aid the estimation of the dependent component of demand in that they can convert demand from one fare group to another fare group [66]. However sell-up rates are generally difficult to obtain as sell-up occurrences are rarely captured during the booking process [6]. The data used to calculate market-specific sell-up rates is inherently sparse. System-wide aggregation methods are often used to make use of more data across the entire system. However estimating sell-up over an entire system results in estimates that are too broad and over-

generalized. This is problematic as estimating sell-up this way may be uncharacteristic of the market of interest. Buy-down behavior is case where customers buy down as a cheaper fare class becomes opened by the airline. Similarly these probabilities are also difficult to calculate in practice for the same reasons. As such, many optimization methods do not directly use sell-up rates to relate demand dependencies between fare classes to maximize revenue.

In order to be able utilize information about customer buying behavior, airlines must first be able to properly quantify these probabilities before determining the availability of various prices. The availability of various prices can be represented by the amount of selling time $\tau_{i,j}$ the airline has allocate to each fare group j of flight i. For each fare group, the demand is dependent on the availability given to it (Figure 1.3). For certain future flights, airlines often employ selling time strategies that are similar to their historical selling times. This is because estimation of demand for scenarios outside of their historical selling time allocations is very difficult. This results in historical data being biased towards these decisions, which implies demand observations are often censored according to the airline's own decisions. This makes the inference of customer buying behavior very difficult, as often-times different fare groups have no common selling time for comparison. Hence, to better understand customer buying behavior, an accurate estimation of demand for each fare group at various selling times is required.

To better estimate demand at common selling times, certain assumptions may be useful for prediction. For example, one common assumption about demand $y_{i,j}(t)$ is that it decreases as price p_j increases. That is, for some common time t, if groups j^- and j^+ relate with prices $p_{j^-} < p_{j^+}$, then the demand relates to each other as $E[y_{i,j^-}(t)] \geq E[y_{i,j^+}(t)]$. Within a fare group, the fact that demand should increase as time increases can also be useful. That is for times t^- and t^+, if the times relate to one another as $t^- < t^+$, then the demand relates to each other as $E[y_{i,j}(t^-)] \leq E[y_{i,j}(t^+)]$. In Chapter 4, these assumptions are mathematically defined via linear operators and bounds. For the assumption on demand decreasing as price increases, Chapter 4 provides a methodology for selecting the lower bound that borrows from the economics formulation of price elasticity.

Building upon these formulations, Chapter 4 also provides a multiple-group Guassian Process formulation of the data such that the constraints can be communicated to the Gaussian Process regression. The idea of implementing these constraints within the constrained Gaussian Process is to allow information to transfer between different fare groups observed at different selling times.

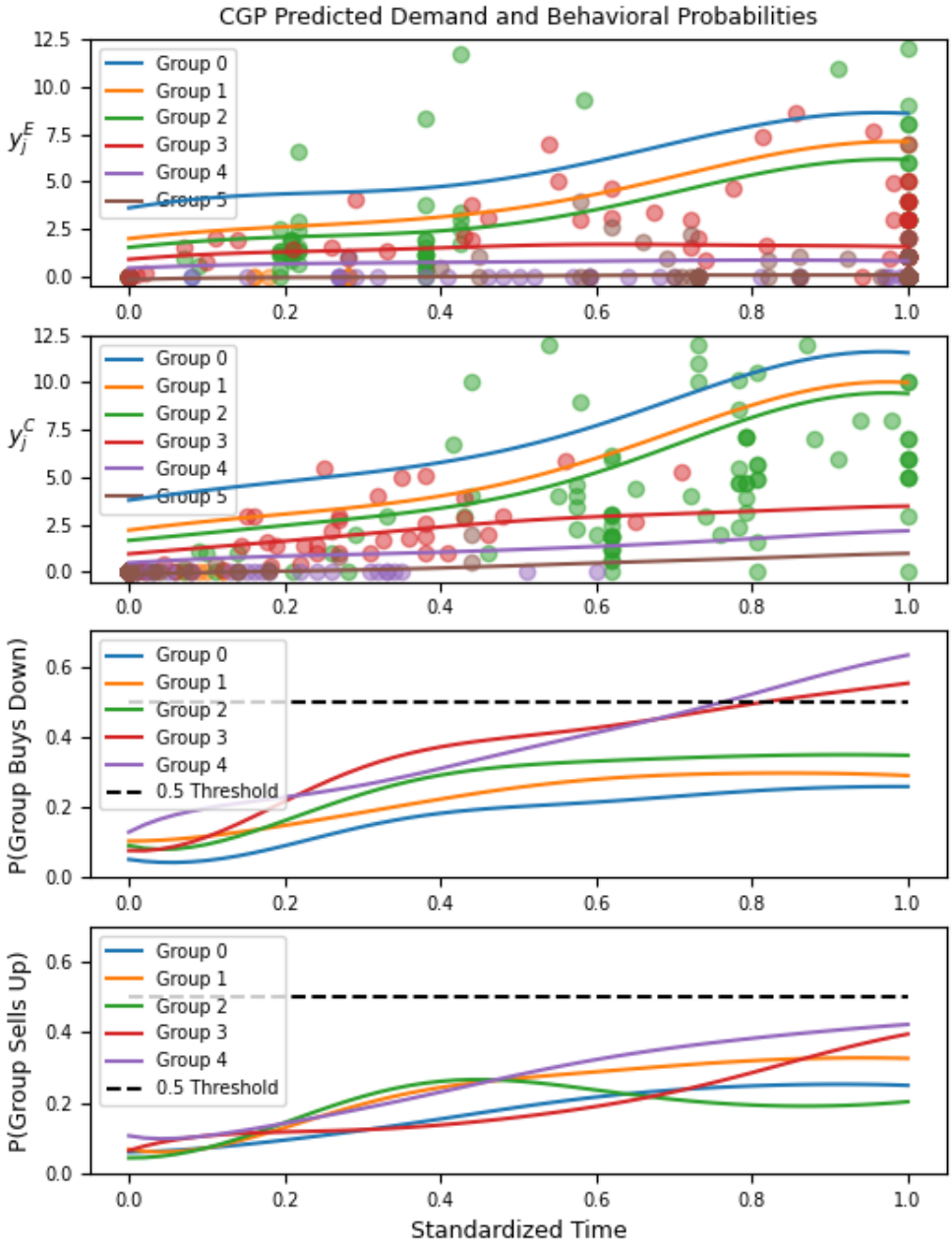


Figure 1.4: *The results of the multi-fare-multi-state constrained Gaussian Process described in Chapter 4. The top two graphs shows the CGP's predictions (indicated by the lines) for data (indicated by the points) belonging to different types of observable demand states for different fare groups. The demand estimates are then converted into buy-down and sell-up probabilities and shown in the bottom two graphs.*

1.2 Key Research Objectives and Contributions

This dissertation is motivated to develop new methodologies that model various types of human behavior. The three main research objectives studied in this dissertation are:

1. Modeling distracted driving behaviors using only kinematic signals belonging to a vehicle. (Chapter 2)

2. Modeling naturalistic human driving behaviors via kinematic data, and ensuring these characterizations are consistent and meaningful. (Chapter 3)

3. Modeling humanistic demand and customer buying behaviors under censored data conditions. (Chapter 4)

The research contributions from each research objective are briefly described in the following subsections.

1.2.1 Contribution from Modeling Distracted Driving Behavior

Chapter 2's contribution lies in a proposed online approach to detect distracted driving behaviors based on the fusion of physics-based kinematic state-space models. The proposed method (shown in Figure 1.5) involves three steps: (1) training multiple kinematic state-space models using historical normal driving data; (2) fusion of the trained motion models for real-time motion prediction via a proposed Autonomous Multiple Model (AMM) algorithm; (3) developing a control chart-based decision strategy by comparing model predictions with the true measurements.

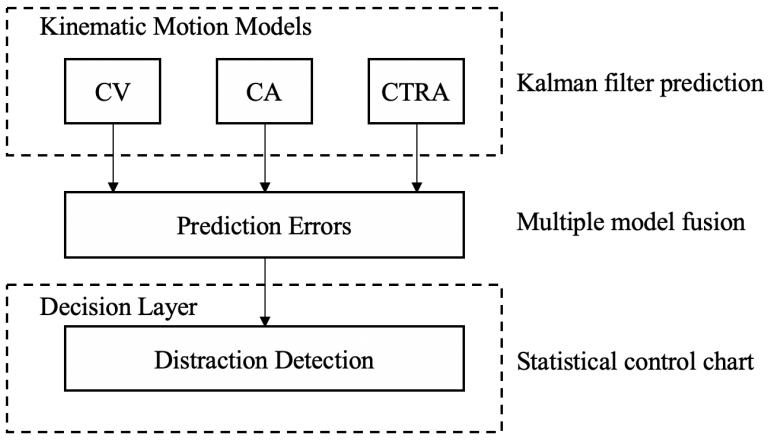


Figure 1.5: *Methodology flowchart for detecting distracted driving (from Chapter 2).*

1.2.2 Contribution from Modeling Naturalistic Human Driving Behavior

Building upon the analysis of kinematic signals, Chapter 3 revolves around the labeling of normal humanistic driving behaviors. Here, an algorithm has been devised to improve the inference behind the Hierarchical Dirichlet Process Hidden Semi Markov Model (HDP-HSMM). The algorithm enhances unsupervised learning by emphasizing important driving behaviors while eliminating redundant ones. The idea is to apply a merging procedure during inference which promotes longer durations and the avoidance of redundant states. In doing so, Chapter 3's contributions will include (1) demonstrating how the HDP-HSMM becomes robust to the inconsistencies brought by the HDP prior; and (2) demonstrating how this paper's method can reduce the number of redundant humanistic driving behaviors shown in Figure 1.1. A brief summary, which describes where the newly described robust HDP-HSMM fits in relation to the other models described in HMM literature, is given in Table 1.1.

State Duration Distribution	**Model**	**Extension (not sensitive to prior)**
Geometric	HDP-HMM [45]	sticky HDP-HMM [12]
Any Discrete Distribution	HDP-HSMM [23]	robust HDP-HSMM (Chapter 3 of this dissertation)

Table 1.1: *Comparison of various HMM-based models versus Chapter 3's proposed robust HDP-HSMM (rHDP-HSMM).*

1.2.3 Contribution from Modeling Humanistic Buying Behaviors

Chapter 4's contribution lies in providing a demand estimation methodology that incorporates airline domain knowledge into the construction of constrained Gaussian Process (CGP) regression. In doing so, the intricate relationship between customer demand for different fare groups under censored data conditions can be modeled to allow for multitask learning to occur. Chapter 4 provides the methodology to facilitate cross-learning via appropriate airline demand constraints, which in turn improves demand estimation. Chapter 4 also takes a unique perspective by incorporating the idea of decomposing demand into different states as to further consider buy-down and sell-up behaviors. Hence a major contribution of this chapter is that customer buy-down and sell-up behavior is probabilistically quantified (Figure 1.4) via the outputs of the described constrained Gaussian Process.

1.3 Outline of Dissertation

The subsequent chapters of this thesis will delve into each work in detail, discussing the context and background of the problems addressed, the methodologies employed, the results obtained, and the implications for the respective domains. Through these endeavors, this dissertation aims to contribute to the advancement of knowledge in human behavior analysis and provide practical applications that can enhance road safety and inform business strategies. The organization of the dissertation is shown in Figure 1.6

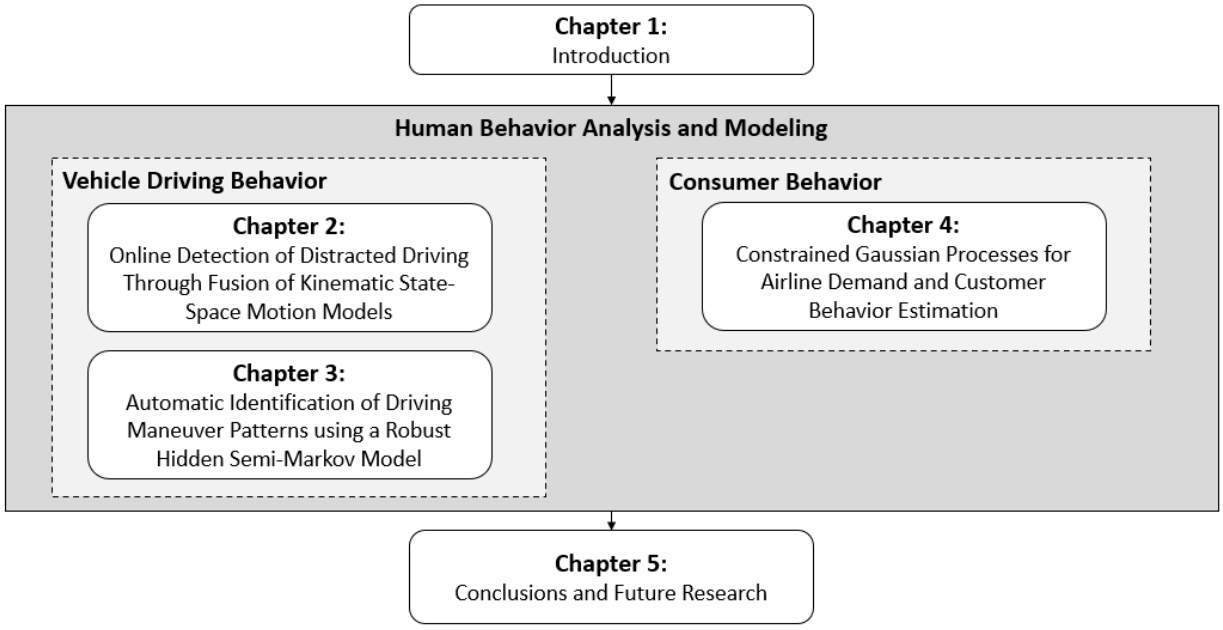


Figure 1.6: *Outline of dissertation.*

CHAPTER 2

Online Detection of Distracted Driving Through Fusion of Kinematic State-Space Motion Models

2.1 Introduction

Driving a vehicle may appear an easy task as it has become almost second nature to today's society. However, in reality driving is a complex decision-making process involving both the driver, their vehicle, and the environment around them. In 2013, the World Health Organization's global status report indicated 1.24 million traffic-related fatalities occurred worldwide and deemed driving to be the current leading cause of death for people aged $15 - 29$ years [40]. Because of statements like this, driving safety is an increasing concern in both transportation research and the automotive industry.

Recently, Driver Assistance Systems (DAS) have been an important focus in the design, development, and manufacturing of new vehicles as not only do they aid drivers in driving, but they also promote safety as well. However there does not yet exist a widely-used DAS capable of recognizing inattentive driving behaviors. This is an important issue because when it comes to vehicular accidents, human failure/error has been implicated to be the cause of about 70% of accidents [61, 1]. From this viewpoint, developing a DAS capable of detecting risky driving behaviors is highly desirable. According to the National Highway Traffic Safety Administration (NHTSA), risky driving behaviors include drunk driving, inattentive driving, not utilizing seat belts, drug impaired driving, drowsy driving. To narrow down the scope further, this paper's focus will be on the issue of inattentive distracted driving.

There is much motivation for developing an algorithm able to detect inattention. For example, in the NHTSA's 100-Car Naturalistic Driving Study, inattention was reported to be 93% of the conflict related to lead-vehicle crashes and minor collisions [8]. Additionally in the United States of 2010, driver inattention was reported to be the direct cause of 3092 deaths and $416,000$ injuries involving accidents [4]. However inattention is also very broadly defined, so this paper will focus on inattention in the form of driver distraction.

Driver distraction occurs when a driver diverts their attention away from the task of driving to focus on a secondary task. For example, texting is a common secondary task performed while driving that causes driver distraction. Much research has been documented on the topic of driver distraction and can usually be categorized into one of the two different approaches.

The first approach focuses on monitoring drivers' behavior directly, as drivers themselves often display certain patterns when becoming distracted. For example, these patterns can be the driver reaching for their phone or even displaying an abnormal blink rate. Research based on the this first approach is usually performed via techniques involving face and facial landmark detection, biomechanical detection and recognition, and measuring cognitive load. [9] provides a great overview on much of the work and successes related to this approach; However, there are many practicalities to consider before the techniques mentioned can be implemented. For example, additional sensors are often required to be either installed in the vehicle or equipped to drivers directly. Not only does incur additional costs to the drivers of interest, but issues privacy may result from implementation. Additionally, sensors such as EEG and ECG electrodes can be intrusive and uncomfortable to drivers. Hence in practice, implementation of above type of research is still very limited.

The other approach focuses on analyzing the vehicle's real-time driving patterns, as any secondary task distracting the driver has a direct influence on his/her body movements that manifest in their control of the vehicle. This approach focuses on using already available motion signals to predict how situations will evolve and measure a driver's deviation from the predicted behavior. The sensors used in this framework are often already embedded in a vehicle measuring information like speed, yaw rate, and acceleration and in turn are more practical than approaches measuring information directly from the driver.

Motion models are often used in parallel to this second approach. They are termed motion models because these models track and predict signals describing the motion of an object. Given this, the structure of a motion model can vary into arguably three different types: (1) data-driven motion models, (2) dynamic motion models, and (3) kinematic motion models. Each structure has its advantages and disadvantages.

Having the most loose of structures, data-driven motion models typically take historical driving data as a training data set to learn typical driving behaviors and maneuvers. For example, [63] used a hierarchical Dirichlet process with a Hidden Semi-Markov Model (HDP-HSMM) to create a set of primitives describing the behavior a particular car given its relative distance and relative range rate to other nearby cars. While data-driven motion models can learn driving patterns directly from historical data, these models typically assume no prior knowledge is known about the number of driving patterns expected to learn. This leads to the problem of learning redundant driving patterns which in turns creates the even greater problem of an incorrect estimation of the duration to each

driving pattern. This can be problematic when referencing these data-driven motion models to check for a driver's deviation to the model's prediction.

On the other hand, dynamic and kinematic motion models make use of the well-defined relationships between the signals of interest defined by physics. This avoids the problem of redundant states and makes the model building process very intuitive. Dynamic motion models are built based on Lagrange's equations and consider the complex physics behind many internal parameters of the vehicle such as longitudinal and lateral tire forces or even the road banking angle.While dynamic models take into consideration the great deal of forces involved in the motion of vehicles, the level of motion prediction required for detecting distracted events can be achieved through the much simpler motion models known as kinematic models.

Kinematic motion models utilize the mathematical relationship between parameters of movement without having to consider the forces that affect the motion. Although much simpler than dynamic models, kinematic models are well known, well established, and can also hand various types of driving assumptions [50]. On the lower end of the scale, linear motion models can be built based on the assumptions of constant velocity (CV) or constant acceleration (CA). These models are simply state-space equations utilizing kinematic equations to capture straight motion without considering rotation. The second level of kinematic models are curvilinear models as they take rotation into account. Assuming no correlation between velocity and yaw rate, curvilinear models include the Constant Turn Rate and Velocity (CTRV) model and the Constant Turn Rate and Acceleration (CTRA) model.

However, not much literature exists relating the use of kinematic motion models to detect driver distraction. One paper that comes close is [57], where four motion models (CV, CA, CTRV, CTRA) with both the extended Kalman filter and particle filter are used to detect lane level irregular driving behaviors via fuzzy logic to output risk types. However their use of GPS signal data presented some challenges in the quality of measurement data. [56] provided an overview on recent papers interested in detecting irregular driving and notes that most of the vehicle-based research is still preliminary and in early stages of development as most do not attempt to quantify performance in terms of correct detection rate nor provide field tests with robust algorithms to classify different types of irregular driving styles. The use of kinematic motion models may be popular in the field of discovering driving primitives, but in the field of detecting driver distraction, most researchers tend to use more visual based approaches.

We propose an online approach to detect distracted driving behaviors based on the fusion of physics-based kinematic state-space models. The proposed method involves three steps: (1) training multiple kinematic state-space models using historical normal driving data; (2) fusion of the trained motion models for real-time motion prediction; (3) developing a control chart-based decision strategy by comparing model predictions with the true measurements. In this section, first

notation, data structure, and the basis of state-space models are introduced. Then, the methodology and preliminary work applying the method to naturalistic driving based on a straight road driving will be presented.

2.2 Methodology Overview

A naturalistic driving dataset includes the measured kinematic signals, which are denoted by the $q-$dimensional vector $y_t = \{y_{1,t}, ..., y_{q,t}\} \in \mathbb{R}^q$. For example, in the IVBSS naturalistic driving dataset, y_t includes the lane offset, yaw rate, longitudinal and lateral velocities and the longitudinal acceleration. At each sample time t, the state variable is denoted as $x_t = \{x_{1,t}, ..., x_{p,t}\} \in \mathbb{R}^p$. Our objective is to develop a statistical model to predict distracted states based on the available measurements $\{y_1, ..., y_t\}$.

The proposed method is divided into three steps. First, the kinematic state-space motion models to be described in Section 2.2.1 are used to predict driving kinematic signals at time $t+1$ given measurements $\{y_1, ..., y_t\}$. Each state-space model represents a different kinematic motion trajectory which is indexed by C_t at sample time t. In this research, three possible state-space models $C = \{CV, CA, CTRA\}$ are used, i.e. $C_t \in C$. Mathematically, the one-step ahead prediction for each model is denoted as $\left(\hat{y}_{t+1} \middle| y_1, ..., y_t; \hat{C}_{t+1|t}\right)$. Second, the state-space models are fused in Section 2.2.2 to make predictions without knowing the actual state-space motion model in advance. To do this, we first estimate the likelihood of each state-space motion model $\mathbb{P}\left[C_{t+1} = c \middle| y_1, ..., y_t\right]$. The prediction is then given by the fused model as:

$$\begin{aligned} \hat{y}_{t+1|t} &= \hat{\mathbb{E}}\left[y_{t+1} \middle| y_1, ..., y_t\right] \\ &= \mathbb{E}_{C_{t+1}}\left[\hat{\mathbb{E}}\left[y_{t+1} \middle| y_1, ..., y_t, C_{t+1}\right]\right] \\ &= \sum_{c \in C} \hat{\mathbb{E}}\left[y_{t+1} \middle| y_1, ..., y_t, C_{t+1} = c\right] \mathbb{P}\left[C_{t+1} = c \middle| y_1, ..., y_t\right]. \end{aligned} \tag{2.1}$$

Third, the prediction errors are calculated as $\epsilon_{t+1} = \hat{y}_{t+1|t} - y_{t+1}$. A control-chart-based decision strategy is developed in Section 2.2.3 to monitor the prediction errors at every real-world second i and hereby label the distraction status $\{\zeta_i\}$. Notice, t reflects 10 Hz frequency updating and i reflects 1 Hz updating. Both a binary label and a distracted level will be provided. An illustrative methodology flowchart describing the steps described is shown in Figure 2.1. Technical details will be discussed as follows.

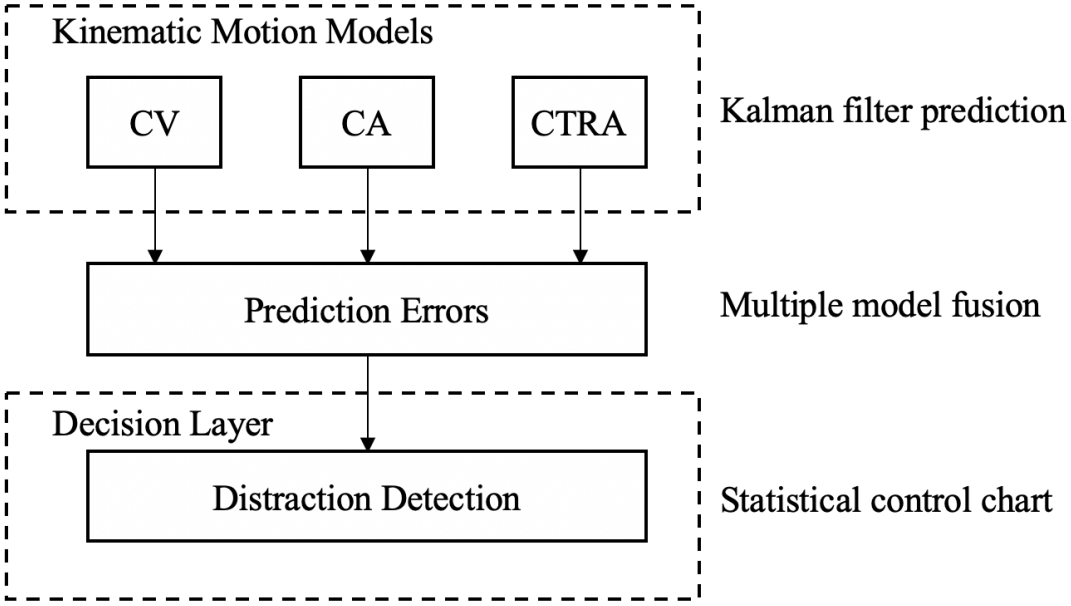


Figure 2.1: *Methodology flowchart. Step 1: Three kinematic motion models (Constant Velocity (CV), Constant Acceleration (CA) and Constant Turn Rate and Acceleration (CTRA) models) are used to predict driving kinematic signals under different scenarios. Step 2: The kinematic motion models are fused to produce a prediction when the scenarios are not known. Step 3: A decision strategy is developed to label* ϵ_t*'s based on the prediction errors* $\hat{y}_{t+1} - y_{t+1}$.

2.2.1 Basis of State-Space Motion Models and Kalman Filter Algorithm

In general, a state-space model specifies the transition function from the driving kinematic state x_t to state x_{t+1} as follows:

$$x_{t+1} = f\left(x_t\right) + w_{t+1},$$

where f is the known transition function depending on the kinematic motion model constraints, and w_{t+1} is the systematic process noise following the multivariate Gaussian distribution $N(0, Q)$. For each time point $t+1$, the measurement y_{t+1} is subject to measurement noise u_{t+1} that follows the multivariate Gaussian distribution $N(0, R)$. The relationship between y_{t+1} and x_{t+1} is determined by the observation equation as follows:

$$y_{t+1} = g\left(x_{t+1}\right) + u_{t+1},$$

where g is the known measurement function. Both Q and R are assumed to be invariant with time. Given this framework, Kalman filters [29] are often used to predict $y_{t+1|t}$ with given measurements $\{y_1, ..., y_t\}$.

When f and g are linear, the state-space model can be written below with F and G denoting the transition and measurement matrices respectively:

$$x_{t+1} = Fx_t + w_{t+1}, \tag{2.2}$$

$$y_{t+1} = Gx_{t+1} + u_{t+1}. \tag{2.3}$$

Under this linear case, the conditional distribution $y_{t+1}\big|y_1, ..., y_t$ is still Gaussian and hence $\hat{y}_{t+1|t}$ has the following closed form:

$$\begin{aligned}\hat{y}_{t+1|t} &= \mathbb{E}\left[y_{t+1}\big|y_1, ..., y_t\right] \\ &= G\mathbb{E}\left[x_{t+1}\big|y_1, ..., y_t\right],\end{aligned}$$

it is sufficient to derive $\mathbb{E}\left[x_{t+1}\big|y_1, ..., y_t\right]$, which is the estimation of the expected kinematic states x_{t+1} given all the historical measurements $\{y_1, ..., y_t\}$. A variety of kinematic motion models have been developed to track vehicle kinematics [50]. The two linear kinematic motion models Constant Velocity (CV) and Constant Acceleration (CA) are employed in this research. Derivations of the Kalman filtering updating procedure in these linear cases can be found in Appendix A.2).

When f or g are not linear, such as in the Constant Turn Rate and Acceleration (CTRA) motion model, the above conditional distribution is analytically intractable. Numerical approaches such as Extended Kalman filter [27], Unscented Kalman filter [28] or Particle filter [48] are usually applied for approximation. In this work, the Extended Kalman filter is used for this nonlinear

case, which approximates the nonlinear f and g using their first order derivatives. This makes the Extended Kalman filter to be easier to realize than the Unscented and Particle filters, which can save computational time in real-time prediction settings.

In the state-space models, R is a diagonal matrix whose diagonal elements are variances of the measurement errors. These values can be obtained from the knowledge on the measurement sensors. The systematic process noise covariance matrix Q describes variability in the difference between the physical model and the actual process. For example, CV model assumes constant velocity during each time interval. Q is related to the error term caused by the constant velocity assumption. In this research, the covariance matrix is estimated based on the driving kinematic signals when the cruise control in engaged during straight road driving. See Appendix A.1 for detailed derivations of f, g, and Q under each state-space motion model (CV, CA, CTRA).

2.2.2 Prediction based on Fusion of State-Space Motion Models

The Kalman filter algorithm predicts the vehicle kinematic signals under the given state-space motion models. Since the state-space motion model C_{t+1} is not known at time t, we would like to fuse the predictions using the expectation over all state-space motion models. Model fusion algorithms have been explored in the existing literature [43]. In this preliminary work, the Autonomous Multiple Model (AMM) algorithm is applied [34] where each Kalman filter updates independently in an autonomous manner.

Suppose, given data $\{y_i\}_{i=1:t}$ and state-space model c, the predicted distributions of state variable x_{t+1} have been generated from the Kalman filter algorithm with mean $\hat{x}_{t+1|t,C_t=c}$ and covariance $P_{t+1|t,C_t=c}$. Denote the prediction of the driving kinematic signals under state-space motion model c by $\hat{y}_{t+1,c}$, and the prediction error by $\epsilon_{t+1,c} = \hat{y}_{t+1,c} - y_{t+1}$. According to the derivation in Appendix A.1, the prediction error $\epsilon_{t+1,c}$ follows a Gaussian distribution $N\left(0, S_{t+1,c}\right)$ with:

$$S_{t+1,c} = R_c + G_c P_{t+1|t,c} G_c'.$$

R_c and G_c are the measurement error covariance and transition matrix of the Kalman filter under state-space motion model c, respectively. The likelihood of prediction errors under state-space model c is estimated as:

$$\begin{aligned}\hat{L}_{t+1,c} &= \mathcal{N}\left(\epsilon_{t+1,c}\middle|0, S_{t+1,c}\right) \\ &= \frac{1}{2\pi|S_{t+1,c}|^{1/2}} \exp\left\{-\frac{1}{2}\epsilon_{t+1,c}' S_{t+1,c}^{-1} \epsilon_{t+1,c}\right\}.\end{aligned}$$

The normalized likelihood $w_{t+1,c}$ is used to weigh the different kinematic motion models' fittings,

resulting in the integrated filtered state:

$$\hat{x}_{t+1|t+1} = \sum_{c=1}^{C} w_{t+1,c}\hat{x}_{t+1|t+1,c},$$

$$w_{t+1,c} = \frac{\hat{L}_{t+1,c}}{\sum_{i\in C}\hat{L}_{t+1,i}},$$

where $\hat{x}_{t+1|t+1,c}$ is the filtered state under state-space motion model c. The resultant $\hat{x}_{t+1|t+1}$ is used to update the Kalman filter in the next iteration. The original weight function:

$$w_{t+1,c} = \frac{w_{t,c}\hat{L}_{t+1,c}}{\sum_{i\in C} w_{t,i}\hat{L}_{t+1,i}}$$

is not recommended because it gives rise to absorbent states while updating - a common case is that $w_{t_1,c} = 0$ leads to $w_{t_2,c} = 0$ for any $t_2 \geq t_1$.

2.2.3 Monitoring Chart Development

In this research, distracted driving behaviors are characterized on the following two data patterns. The first is unstable driving patterns reflected by variance of kinematic variables. As lane offset is a sensitive indicator of driving performance, a Cumulative Sum control chart (CUSUM chart, [41]) is generated to monitor the variance of the lane offset's prediction error. The second is large deviations of lane offset, which is monitored by an Exponentially Weighted Moving Average control chart (EWMA chart, [37]). These two control charts are combined to provide a multi-level assessment of distraction severity.

In this research, the kinematic data is sampled at a 0.1 second interval, but the decision is made at a 1 second interval. Thus, the two control charts are established on non-overlap time windows of $i = 1, 2, 3, \ldots$, where each window contains $n = 10$ time points. Let $\epsilon_{t,\text{CTRA,L}}$ represents the lane offset (denoted by L) prediction error at time t under state-space motion model CTRA. To monitor the error variance, a standardized quantity v_i [20] is computed for each time window i as:

$$v_i = \frac{\sqrt{|(\bar{\epsilon}_i - \mu_0)/\sigma_0|} - 0.822}{0.349},$$

$$\bar{\epsilon}_i = \frac{1}{\sqrt{n}} \sum_{t=n(i-1)}^{ni-1} \epsilon_{t,\text{CTRA,L}},$$

where μ_0 and σ_0 are normalization parameters that can be estimated as the sample mean and standard deviation of lane positions taken from data in CTRA scenarios. Since the in-control

distribution of v_i is approximately $N(0,1)$, the one-sided standardized scale CUSUM chart can be established as follows:

$$S_i^+ = \max\left[0, v_i - k + S_{i-1}^+\right],$$

where initial value $S_0^+ = 0$, and k reference/allowance/slack value for the CUSUM statistic (often chosen to be about half of a shift in standard units). The CUSUM statistic S_i^+ is sensitive to variance changes of v_i, and increases if the standard deviation of v_i increases. As such, the CUSUM chart is alarmed at times i whenever $S_i^+ > h$, where h represents the decision threshold of the CUSUM chart.

The EWMA chart is selected to monitor small mean shifts of lane offsets since it is more robust for non-Gaussian variables. Suppose $x_{t,\mathrm{L}}$ is the lane position at time t. For a given window size n and a constant λ, the monitoring statistics $\tilde{x}_i$ at window i are updated recursively as:

$$\tilde{x}_i = \lambda \bar{x}_i + (1-\lambda)\tilde{x}_{i-1},$$

where $\bar{x}_i = \frac{1}{n}\sum_{t=ni}^{n(i+1)-1} x_{t,\mathrm{L}}$ and $\tilde{x}_0 = \bar{x}_1$. The L-width control limits are calculated as:

$$UCL_i = L\sigma_1\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]},$$

$$LCL_i = -L\sigma_1\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]},$$

where σ_1 is the standard deviation of lane positions during straight road driving. The EWMA chart alarms whenever $\tilde{x}_i > UCL_i$ or $\tilde{x}_i < LCL_i$, alarming only when small shifts in the mean ($\leq 1.5\sigma_1$) are detected.

By combining the CUSUM and EWMA monitoring results, the distraction is evaluated with the following five levels (Level 1-5) to reflect the severity of distractions from low to high. The in-vehicle alarm systems will be designed to indicate different levels of distractions.

- Level 1: The EWMA chart generates an alarm at window i;
- Level 2: The CUSUM chart generates an alarm at window i;
- Level 3: Both the two control charts generate alarms at window i;
- Level 4: The two control charts generate alarms at two consecutive windows i, $i-1$;
- Level 5: The two control charts generate alarms at three consecutive windows i, $i-1$, $i-2$.

To avoid continuous alarms due to cumulative large variances in the CUSUM chart, the monitoring statistic S_i^+ is reset to zero whenever a Level 5 alarm is generated.

2.3 Case Study

In this section, the effectiveness of the proposed method is demonstrated in several naturalistic straight-road driving datasets. The model parameters are trained based on 22 non-distracted, naturalistic, straight-road driving events selected from the IVBSS dataset. In the training set, each time point is assigned with a binary distraction label where $\zeta_t = 0$ means the driver is not distracted and $\zeta_t = 1$ otherwise. The distraction status is manually labelled according to the videos recorded by the in-vehicle cameras. The trained online detection algorithm is then applied to a different testing dataset. The detection rates and false alarm rates are calculated for evaluation purpose. In the datasets, driving signals including velocity, lane position and yaw rate are measured with a sampling frequency of 10 Hz.

In the algorithm, the R matrix is built based on the standard deviations of velocity's, lane position's, and yaw rate's measurement errors, and were set as 0.45 m/s, 0.01 m and 0.1 degree/s according to the domain knowledge. The values of trained parameters are listed as follows. The Q matrix is built based on the standard deviations of longitudinal acceleration, lateral acceleration and angle acceleration, which were estimated as 0.3 m/s^2, 0.1 m/s^2 and 0.1 $degree/s^2$ in the training set. The μ_0 and σ_0 in the CUSUM chart are set as 0 and 0.0523 m, respectively.

Seven events are selected from the IVBSS datasets to evaluate the detection performance, each of which has a cell phone use (texting) period. Detection results for two of the selected events are illustrated in Figures 2.2 and 2.3. Alarms are generated only for Level 5 distractions (shown in the fifth row of both figures). Figure 2.2 shows a case of misdetection. Even though the driver was using a cell phone, the lane position signals are still stable, implying that the driver is good at texting while driving. In this case, the distraction behavior is not considered to be risky. However, in the Figure 2.3, an alarm is generated during the texting period.

The algorithm is also applied to two other datasets collected from the IVBSS and the SPMD dataset where no distraction happened to test the false alarm rate. The detection rate is calculated as the frequency of the detected events divided by the total number of the events. The false alarm rate (where alarms occur at Level 5 only) is evaluated as the number of false alarms divided by the total length of time. A summary of the results in all the four datasets is shown in Table 2.1.

The only misdetection (also shown in Figure 2.2) can be explained by the good driving skill of the distracted driver. The videos recorded by the in-vehicle camera are checked to investigate the causes of false alarms. One of the main causes is passing vehicles, especially heavy trucks. In fact, a driver deviates from the lane center and accelerates when trying to pass another vehicle, which results in increments in variabilities of both lane positions and velocities. These false alarms can be avoided by either recognizing the vehicle patterns based on data or installing a sensor to detect surrounding vehicle in the future work.

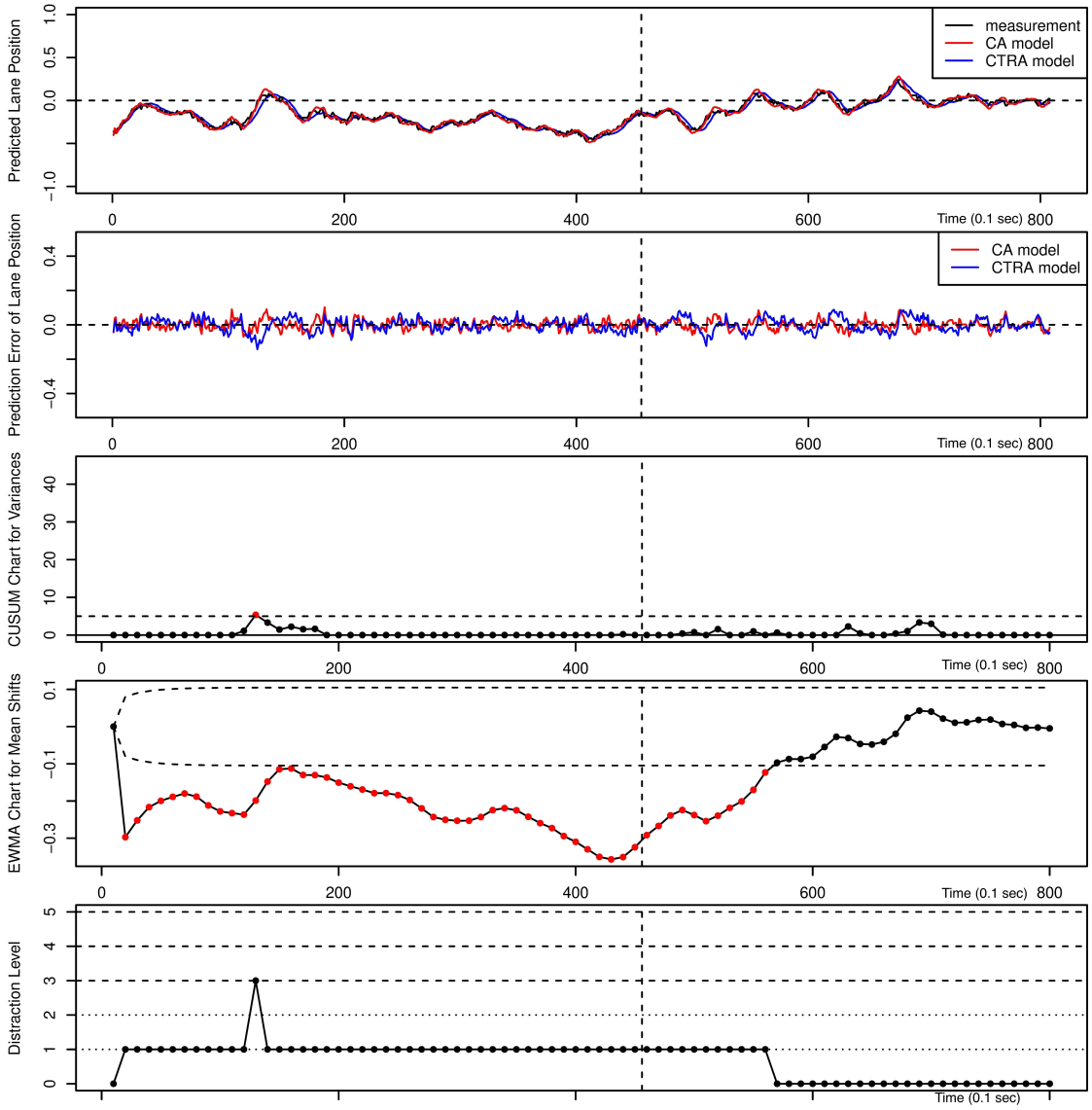


Figure 2.2: *Detection results based on the CUSUM (third row) and EWMA (fourth row) monitoring results. Alarms are only generated when a Level 5 (shown in the fifth row) distraction occurs. As the vertical dashed line represents the start time of the true distraction, it can be observed that the algorithm misses detecting this distracted event. This is believed to have been caused by the small prediction error (second row).*

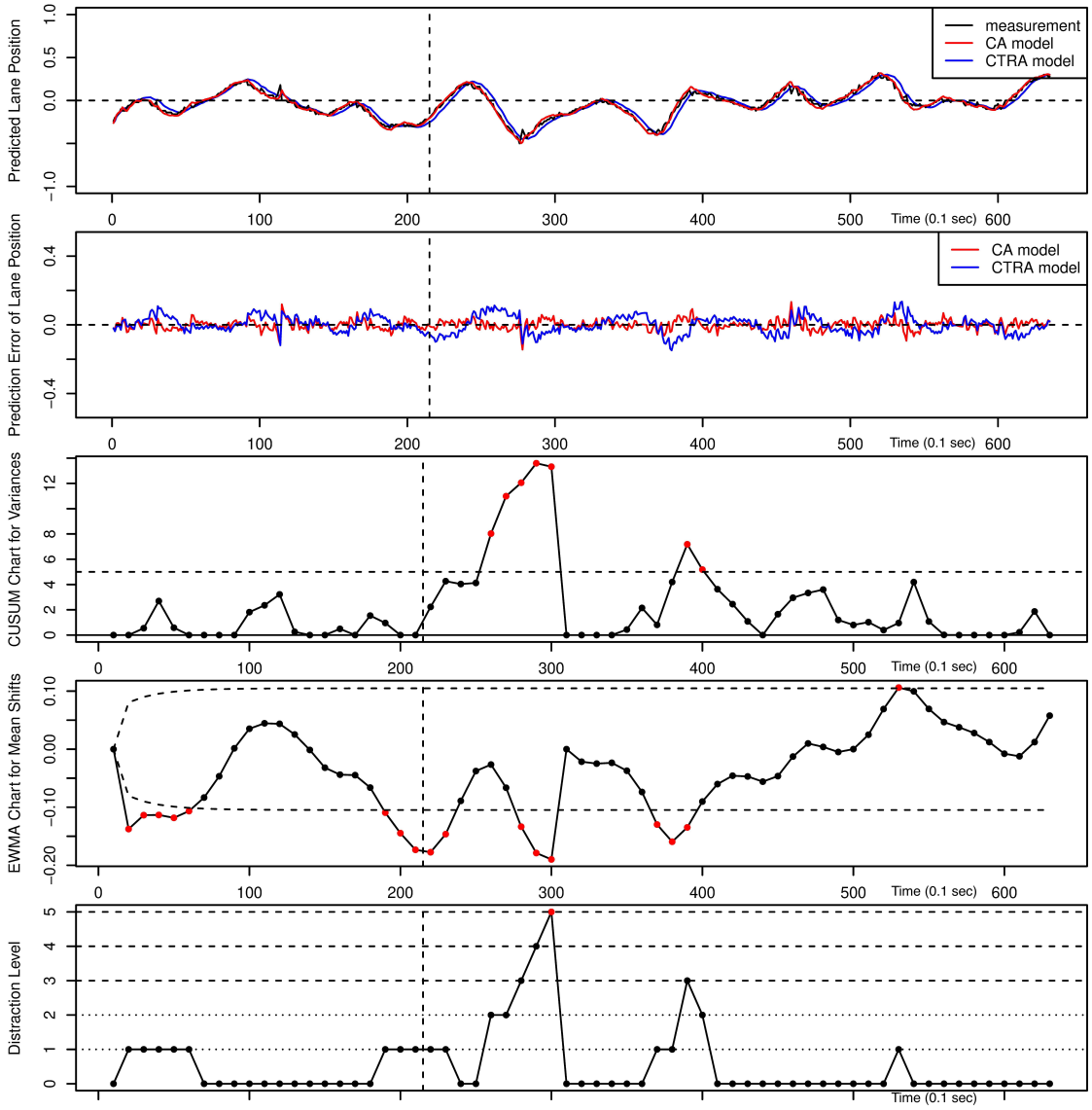


Figure 2.3: *Detection results based on the CUSUM and EWMA monitoring results. Alarms are only generated when a Level 5 distraction occurs. The vertical dashed line represents the start time of the true distraction. In this case, the algorithm detects this distraction 9 seconds past the start of detection, with the distraction level increasing up to the time of detection.*

Table 2.1: *Summarized Results*

#	Dataset	Include Distraction	Include Curvature	Include Lane Changes	Detection Rate	False alarm Rate
1	IVBSS (training)	No	No	No	NA	0/692s
2	IVBSS	Yes	No	Yes	6/7	3/470s
3	IVBSS	No	Yes	Yes	NA	47/3880s
4	SPMD	No	No	No	NA	0/762s

2.4 Discussion

In summary, an online detection algorithm is developed to monitor distractions under the straight driving kinematic signals. To predict vehicle kinematics, three typical kinematic motion models are integrated. In the decision layer, two monitoring charts (CUSUM and EWMA) are combined to provide a multiple level assessment of distraction severity.

In this preliminary work, the algorithm is only applicable for detecting distractions during straight road driving on highways. To extend the algorithm for curvy road driving, CTRA state-space models has already been developed. The next step for improvement is to use the HDP-HSMM to estimate transition probabilities for motion model fusion, and to further use the HDP-HSMM to handle other driving maneuvers when the state-space models are not applicable. The details of building this HDP-HSMM will be discussed in Chapter 3.

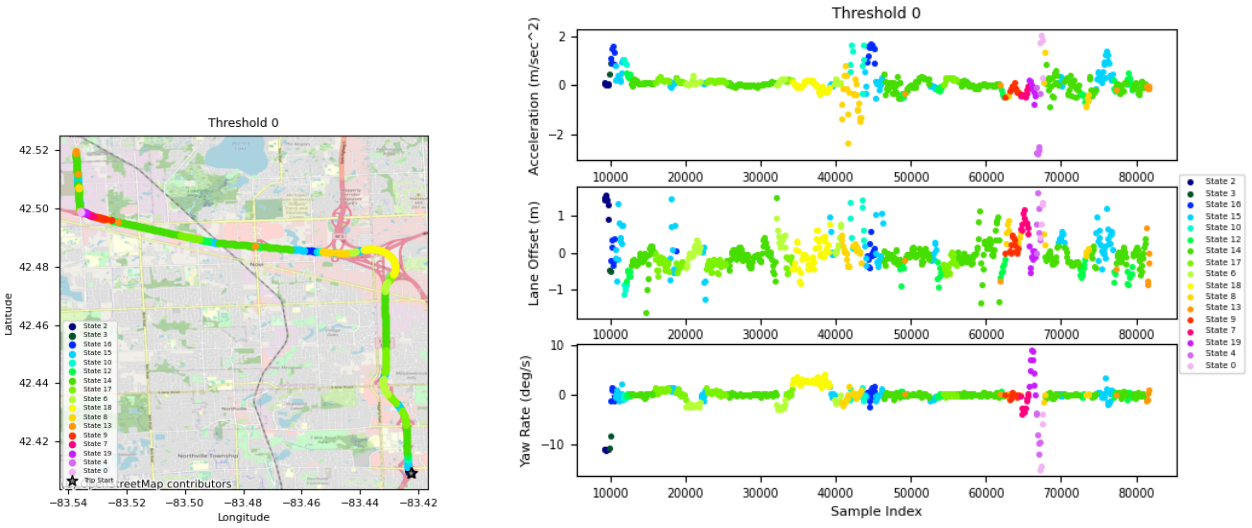
CHAPTER 3

Automatic Identification of Driving Maneuver Patterns using a Robust Hidden Semi-Markov Models

3.1 Introduction

The analysis of vehicle driving styles is prominent to the field of intelligent transportation and vehicle calibration [69, 46]. The term *driving style* can be referred as a set of dynamic activities or steps that a driver uses when driving. Hence, this type of research impacts eco-driving, road safety, and intelligent vehicles [7, 49, 35]. To model these driving styles, one popular approach is the use of a Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM) [62]. This model is powerful in that it considers the sequential nature of driving kinematic signals, and estimates data segmentation, behavior state duration, and state transition probabilities. The HDP-HSMM provides semantical way for analyzing driver behaviors, and is thus popularly used for describing driving styles. Figure 3.1b shows an exemplar set of sequential kinematic signals belonging to the trip observed in Figure 3.1a. The signals are color-coded to reference a state segmentation determined by a HDP-HSMM.

While the HDP-HSMM is powerful, literature outside of the field of transportation details how the model's use of an HDP prior can lead to redundant and inconsistent state estimations. This detail is important as it needs to be considered by researchers attempting to utilize the HDP-HSMM to describe driving styles. For example, Figure 3.1 clearly has redundant states as seen by the green shaded states. The redundant states can make analysis of HDP-HSMM outputs across multiple datasets difficult for researchers hoping to utilize the HDP-HSMM to model driving styles. This paper addresses this issue by presenting an algorithm that reduces redundant states to improve consistency while still aligning to the structure of a basic HDP-HSMM. The presented algorithm results a more robust HDP-HSMM (rHDP-HSMM) that is expected to output a more consistent data segmentation, behavior state duration, and state transition probabilities than a basic HDP-HSMM.



(a) (b)

Figure 3.1: *An example trip and the kinematic signals belonging to it. Learned states from an HDP-HSMM are color coded as labels.*

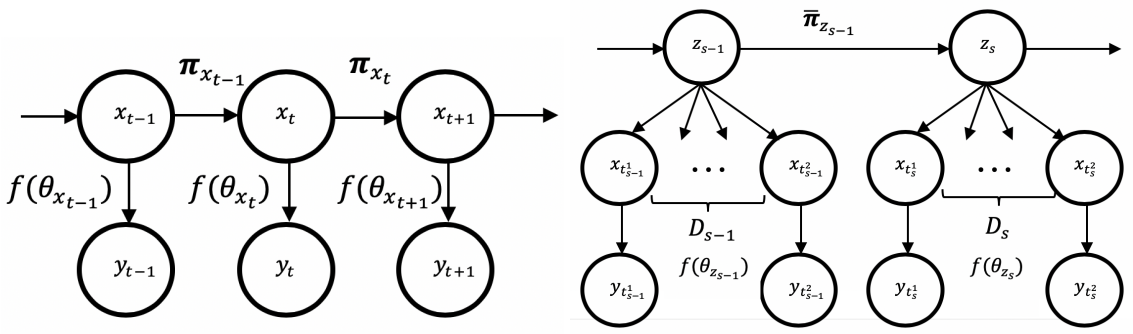
This will impact the transportation field in that driving maneuver patterns can be better grouped together for classification or behavioral studies.

The remainder of this paper is as follows. Section 3.2 will provide the background about HDP-HSMM's from a statistical perspective, and highlight the current set of approaches towards addressing the issues derived from the HDP prior. Section 3.3 will provide the data description and the model formulation of a basic HDP-HSMM. Section 3.4 discusses the details of inference for a HDP-HSMM, and how this paper's algorithm can be included within the inference to produce a more robust HDP-HSMM. Section 3.5 presents a simulation study, in which the rHDP-HSMM is compared to the basic HDP-HSMM based on simulated data. Section 3.6 presents a case study that uses realistic, naturalistic driving data to compare the rHDP-HSMM with the original HDP-HSMM method on the basis of describing driving patterns. Finally, Section 3.7 summarizes new contributions and major conclusions of the paper.

3.2 Background

The HDP-HSMM was designed to improve upon the structure of a discrete state-space Hidden Markov Model (HMM). HMM's are also popularly used for describing sequential data [12, 67, 33, 32, 58, 30]. In particular, the HMM [53, 45] utilizes a two-layer structure (Figure 3.2a) to represent sequential data observed at equally spaced time points. In this model, data is assumed to

be generated from a set of probability distribution functions dependent on corresponding hidden states. The hidden states determine the data segmentation. Transitions among hidden states are modeled as a Markov Chain. This allows for the consideration of time sequence information during inference and further aids in the prediction of future states. One condition of using the Markov Chain is that the state duration of each hidden state is assumed to be Geometrically distributed.



(a) *HMM* **(b)** *HSMM*

Figure 3.2: *A comparison between the structure of a Hidden Markov Model (HMM) and a Hidden Semi-Markov Model (HSMM). The variables and their descriptions are as follows: x_t (hidden state at time t), y_t (observed data at time t), π_x (transition probabilities of state x), $f(\theta_x)$ (probability distribution of state x), z_s (state of segment s), D_s (state duration of segment s).*

While the HMM is able to define data segmentation and state transitions, its definition of state duration is severely limited by the model's structure. This limitation lead to the development of the Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM) [23] which provided two key improvements to the HMM. The first improvement was the removal of the HMM's assumption of geometrically distributed state duration. As the HDP-HSMM uses a Semi-Markovian approach to model the state transitions $\bar{\pi}_{z_s}$, this removes self-transitions from the transition matrix. As a consequence, this frees the geometric distribution restriction on the duration D_s, which leads to a three-layer structure model as shown in (Figure 3.2b). In other words, users can choose different models for representing state duration, while allowing the segmentation of hidden states to be directly represented by z_s.

The second improvement was the introduction of Dirichlet Processes to the model. The Dirichlet processes is an extension to the Dirichlet distribution, as atoms can be sampled from it based on an input distribution. However, one key difference is that the Dirichlet Process assigns a probability of drawing a new atom from the input distribution and a separate probability of drawing an atom based on the atoms seen in previous samples. The resulting distribution is discrete and similar to the input distribution, but also has the possibility of having infinite discrete atoms if infinite samples were drawn. This phenomenon is interesting in the context of HMMs and HSMMs, as the Dirichlet

process can be used as a prior to the state transition probability vector [3, 60, 23]. Doing this allows the probability vector length (i.e. models' number of states) to grow without limit during inference, which implies the Dirichlet process also acts like a prior on the number of clusters. In the HDP-HSMM, a Hierarchical Dirichlet Process (HDP) is used as a prior on the state transitions, which allows all the state transition probabilities to share a similar base distribution. This is beneficial, as all the states represented in the base distribution are shared between all the different state transition probabilities, while allowing each transition probability be dependent on the exit state. Hence, for the context of modeling of driving maneuvers, the HDP-HSMM is preferred as it allows greater flexibility in defining the relationship between the data and segmentation, state duration, and state transitions.

While the Dirichlet Process's clustering properties have been seen as a tool to address the model selection for Bayesian nonparametric approaches [59, 25], the Dirichlet Process is known to have inconsistency issues regarding estimation of the true number of states. [36] provided an example for Dirichlet Process Mixture Models which demonstrates how the posterior does not concentrate at the true number of components, and instead introduces extra clusters even if they are not needed. Under the context of HMMs, [26] showed how the Dirichlet Process also leads to the creation of redundant states, which presents an unrealistic rapid switching between states in the inferred transition matrices. Under the context of HSMM's, Figure 3.1 shows how this side effect occurs even in the HDP-HSMM. However, for the HDP-HSMM, the redundancy issue also affects the inference of transition probabilities and duration estimation.

A few works exist that focus on solving this issue for HMM's. [15] discussed HMM's utilizing a Dirichlet prior, and the assumptions on the prior required for the consistency. [11] developed the *sticky* HDP-HMM (sHDP-HMM) to consider the issue of redundant states. This model adds a bias to the prior on the rows of the transition matrix which emphasizes self-transitions. This results in an increased state duration for each learnt state, which allows the sHDP-HMM to avoid redundant states with short state duration. However, this strategy cannot be applied to HDP-HSMM as the modeling structure of HMM's is inherently different from HSMM's. Outside of HMM and HSMM modeling, [18] focused on the Dirichlet Process Mixture model, and presented the Merge-Truncate-Merge algorithm, which guaranteed a consistent estimate to the number of mixture components. This post-processing procedure takes advantage of the fact that the posterior sample tends to produce a large number of atoms with small weights, and probabilistically merges atoms together.

Given these approaches, this paper attempts to address the HDP's inconsistency problem by taking inspiration from both the sticky HDP-HMM and the Merge-Truncate-Merge algorithm. The idea is to apply a merging procedure during inference which promotes longer durations and the avoidance of redundant states. In doing so, this paper's contribution will include demonstrating how the HDP-HSMM becomes robust to the inconsistencies brought by the HDP prior and how

this paper's method can reduce the number of redundant states to better define driving maneuvers existing in Figure 3.1a. A brief summary, which describes where our model fits in relation to the other models described in HMM literature, is given in Table 3.1.

State Duration Distribution	Model	Extension (not sensitive to prior)
Geometric	HDP-HMM [45]	sticky HDP-HMM [12]
Any Discrete Distribution	HDP-HSMM [23]	robust HDP-HSMM (This paper)

Table 3.1: *Comparison of various HMM-based models versus our proposed robust HDP-HSMM (rHDP-HSMM).*

3.3 Problem Formulation

3.3.1 Data Description

In this paper, a sequential dataset consists of a series of observations collected at T chronologically ordered time points. At each time point t, $y_t \in \mathbb{R}^p$ represents the $p-$dimensional signal responses. The sequential data is assumed to follow multiple phases; there exists a partition $1 = t_1^1 \leq t_2^1 \leq ... \leq t_S^1 = T - D_S$, such that the elements within the $s-$th segment, denoted by $y_{t_s^1:t_s^2}$, are independent and identically distributed (i.i.d.) for a state duration of $D_s \in 1, 2, \ldots, S$.

The objective of the data analysis is generalized to (1) identify distributional patterns that describe each phase, (2) identify the time duration distribution corresponding to each segment, and (3) identify the probability of transitioning from one distribution to another. The challenge lies in little information being available relating to the number of states, the states' durations, and the transition probability matrix.

3.3.2 Basis of HDP-HSMMs and Notations

The HDP-HSMM accomplishes this objective with the following structure. The multivariate sequential data is represented by the sequence $(y_t)_{t=1:T} := \{y_t \in \mathbb{R}^p : t = 1, ..., T\}$ and is assumed to transit among K different hidden states. The hidden states at each time point t are represented by the sequence $(x_t)_{t=1:T} := \{x_t \in \{1, 2, \ldots, K\} : t = 1, ..., T\}$, and can be further divided into S segments. Within each data segment $s \in \{1, 2, \ldots, S\}$, all hidden states share the same index (labeled by the super-state $z_s \in \{1, 2, \ldots, K\}$), and the state duration of the segment is denoted by D_s. As such, the start and end times of each segment s are indexed by time stamps t_s^1 and t_s^2,

respectively. They can be calculated as $t_s^1 = \sum_{\bar{s}<s} D_{\bar{s}}$ and $t_s^2 = t_s^1 + D_s - 1$ where $\bar{s}$ represents all the segments before segment s. The state of segment s is assumed to be Markovian with a transition probability $\pi_{i,j} = \Pr(z_s = j \mid z_{s-1} = i)$, where the rows of the transition matrix are denoted as $\pi_i = [\pi_{i,1}\ \pi_{i,2}\ \ldots \pi_{i,K}]$. However, as each state has a random state duration $D_s \sim g(\omega_{z_s})$, the HSMM does not permit self-transitions to occur. To consider this, the transition rows of π_i are adjusted to $\bar{\pi}_i$ with each element being $\bar{\pi}_{i,j} = \frac{\pi_{i,j}}{1-\pi_{i,i}}(1 - \delta_{i,j})$ (where $\delta_{i,j} = 1$ if $i = j$; $\delta_{ij} = 0$ otherwise).

The relationship between the observation sequence and the segmentation described above can be seen by the emission distribution functions $f(\theta_{z_s})$ and the state duration probability mass functions $g(\omega_{z_s})$ with parameters θ_{z_s} and ω_{z_s} being dependent on segment s. The priors on θ_{z_s} and ω_{z_s} are denoted by H and G respectively.

A Hierarchical Dirichlet Process (HDP) is used to define a prior on the rows of the transition matrix (π_i) to learn the number of unknown states. The HDP creates a countably infinite state-space and utilizes a stick-breaking process $\beta \sim \text{Beta}(\gamma)$ [51] to determine the number of unknown states (K). A smaller γ ($\gamma \geq 0$) yields more concentrated distributions, which plays a part in shaping the transition pattern. Each row of the Markovian transition probability matrix is sampled from a Dirichlet process ($\pi_i \overset{\text{iid}}{\sim} \text{DP}(\alpha, \beta)$) and its similarity to the stick-breaking process depends on the concentration parameter $\alpha \in (0, \infty)$.

The HDP-HSMM is shown in Figure 3.2b and can be formulated as follows:

$$
\begin{aligned}
&\beta \sim \text{Beta}(\gamma), \\
&\pi_i \overset{\text{iid}}{\sim} \text{DP}(\alpha, \beta) \quad (\theta_i, \omega_i) \overset{\text{iid}}{\sim} H \times G \qquad i = 1, 2, \ldots, \\
&\quad z_s \sim \bar{\pi}_{z_{s-1}} \\
&\quad D_s \sim g(\omega_{z_s}) \qquad\qquad\qquad\qquad s = 1, 2, \ldots, \\
&\quad x_{t_s^1:t_s^2} = z_s, \\
&y_{t_s^1:t_s^2} \overset{\text{iid}}{\sim} f(\theta_{z_s}) \qquad t_s^1 = \textstyle\sum_{\bar{s}<s} D_{\bar{s}} \qquad t_s^2 = t_s^1 + D_s - 1.
\end{aligned}
\tag{3.1}
$$

Typically, Gibbs sampling approaches are used for statistical inference of the model parameters of the HDP-HSMM, which requires the full conditional distributions of the model parameters [16]. The details of the general Gibbs sampling procedure and how this paper applies a merging algorithm within it to create a robust HDP-HSMM is presented in the next section.

3.4 Proposed Robust HDP-HSMM

3.4.1 Inference

The details of the block sampling procedure presented in [23] to infer the parameters for the HDP-HSMM are discussed here. Additional insight regarding this paper's proposed changes will also be included in this section. Assume initial values have been set for the state sequence, the emission parameters, the duration parameters, and the transition probabilities:

$$(x_t)^{(0)}, \{\theta_i\}^{(0)}, \{\omega_i\}^{(0)}, \{\pi_i\}^{(0)}.$$

Step 1: The block sampling procedure begins iteration $m = 1$ with the sampling of the emission, duration, and transition distribution parameters. The distributional parameters can be sampled independently of one another, conditional on data assigned to each state i under the current state sequence $(x_t)^{(m-1)}$. Assuming distributions with conjugate priors are utilized within the HDP-HSMM, this step can be simplified significantly into the following statement:

$$\begin{aligned}
\{\theta_i\}^{(m)} &\sim h_{\theta_i}(\theta_i|(x_t)^{(m-1)}, (y_t), H, G, \beta) \\
\{\omega_i\}^{(m)} &\sim h_{\omega_i}(\omega_i|(x_t)^{(m-1)}, (y_t), H, G, \beta) \\
\{\pi_i\}^{(m)} &\sim h_{\pi_i}(\pi_i|(x_t)^{(m-1)}, (y_t), H, G, \beta),
\end{aligned}$$

where h_θ refers to the updated posterior corresponding to the conditional distribution with parameter θ.

Step 2: Once a new set of parameters have been sampled, it is practical to apply some identifiability constraints to the parameters to help ensure state switching does not occur during the sampling procedure. State switching is a problem mentioned in literature [22, 54], in which the permutation of defined states is not considered during the sampling procedure. Identifiability constraints ensure the order of states does not change between iterations of the sampling procedure, and helps ensure the posterior chain is not multimodal at the end of the sampling procedure. While many types of constraints can be applied, such as rearranging the states such that $\theta_1 < \theta_2 < \theta_3 < \ldots$, the constraints used in this paper are be mentioned in each section directly.

Step 3: After identifiability constraints have been applied, the new state sequence can be

sampled. [23]'s procedure makes use of the following backwards messages:

$$\begin{aligned}
B_t(i) :=& p(y_{t+1:T}|x_t = i, F_t = 1) \\
=& \sum_j B_t^*(j) p(x_{t+1} = j|x_t = i) \\
B_t^*(i) :=& p(y_{t+1:T}|x_{t+1} = i, F_t = 1) \\
=& \sum_{d=1}^{T-t} B_{t+d}(i) p(D_{t+1} = d|x_t = i) p(y_{t+1:t+d}|x_{t+1} = i, D_{t+1} = d) \\
&+ p(D_{t+1} > T - t|x_{t+1} = i) p(y_{t+1:T}|x_{t+1} = i, D_{t+1} > T - t) \\
B_T(i) :=& 1,
\end{aligned}$$

where $F_t = 1$ denotes a new segment begins at $t + 1$, and D_{t+1} denotes the duration of the segment that begins at time $t + 1$ [38]. The procedure for obtaining the posterior state sequence begins by drawing a sample for the first state using the following formula:

$$p(x_1 = k|y_{1:T}) \propto p(x_1 = k) B_0^*(k).$$

Next, a sample is drawn from the posterior duration distribution by conditioning on sampled initial state $\bar{x}_1$:

$$p(D_1 = d|y_{1:T}, x_1 = \bar{x}_1, F_0 = 1) = \frac{p(D_1 = d) p(y_1 : d|D_1 = d, x_1 = \bar{x}_1, F_0 = 1) B_d(\bar{x}_1)}{B_0^*(\bar{x}_1)}.$$

The rest of the state sequence can be sampled assuming the new initial state has distribution $p(x_{D_1+1} = i|x_1 = \bar{x}_1)$ and repeating the process, until a state is assigned for all indices $t = 1, \ldots, T$.

Step 4: Once the new state sequence is sampled, the Gibbs sampling procedure normally returns back to Step 1, increments m by 1, and repeats Steps 1 to 3 until posterior convergence. However, before doing that, this paper propose adding an additional sampling Step 4 that removes redundant states from the posterior state sequence

$$(\tilde{\beta}, \tilde{x}_t)^{(m)} \sim h_{(\tilde{x}_t)}((\tilde{x}_t)|\{\theta_i\}^{(m)}, \{\omega_i\}^{(m)}, \{\pi_i\}^{(m)}(x_t)^{(m)}, (y_t), H, G, \beta), \tag{3.2}$$

where $h_{(\tilde{x}_t)}(\cdot)$ represents a sampling step proposed by this paper to promote robustness.

3.4.2 Implementation of Step 4

The proposed Step 4 is the main contribution of this paper. This section will provide the details on how to implement Equation 3.2 described in Step 4 above. The procedure is described by first

defining redundancy between two states:

Definition 3.4.1 *In the state sequence $(x_t)_{t=1:T}$, the states i and j are identified as* ***redundant states*** *if $\mathcal{D}\big(f(\theta_i), f(\theta_j)\big) \leq \tau$, where τ is the decision threshold and $\mathcal{D}\big(f(\theta_i), f(\theta_j)\big)$ is a measure of divergence that gets larger when the distributions $f(\theta_i)$ and $f(\theta_j)$ are more different from one another.*

Although $\mathcal{D}\big(f(\theta_i), f(\theta_j)\big)$ can be any measure of divergence satisfying Definition 3.4.1, the remainder of the paper will assume $\mathcal{D}\big(f(\theta_i), f(\theta_j)\big) = ||(\theta_i - \theta_j)||_2$ is the ℓ^2 norm of the difference in parameters.

Now that redundancy has been defined, the details of Equation 3.2 can be represented by Algorithm 1. In short, the procedure samples a new state sequence that contains no redundant states. [23] describes a weak-limit approximation to the Dirichlet Process prior,

$$\begin{aligned} \beta|\gamma &\sim \text{Dir}(\gamma, \ldots, \gamma) \\ \pi_j|\beta &\sim \text{Dir}(\alpha\beta_1, \ldots, \alpha\beta_K), \quad j = 1, \ldots, K, \end{aligned}$$

as well as an augmentation that introduces auxiliary variables which are added to the β vector to preserve conjugacy. This approximation eases the use of sampling procedures when dealing Dirichlet Processes [24]. Taking this approach, the β vector takes no consideration of redundant states, which may negatively impact the posterior of π_j. The presence of redundant states means the posterior transition probabilities contain extra transitions to and from redundant states, which dilute the underlying transition process. To counter this, $h_{(\tilde{x}_t)}(\cdot)$ aims to adjust the β vector in this step as to discourage transitions to redundant states in future steps, and preserve the true underlying transition process.

Algorithm 1 describes $h_{(\tilde{x}_t)}(\cdot)$ entirely. The procedure begins by initializing a new vector $\tilde{\beta}$, a new state sequence $(\tilde{x}_t)^{(m)}$, and taking the input of a similarity threshold τ. Taking inspiration from [18], the states order is firstly randomized in which redundancy is checked. This is to ensure the start of the merging procedure begins at a point close to the "central mass" of the emission distribution clusters with a high probability. Going through the order, if the state exists within the new state sequence $(\tilde{x}_t)^{(m)}$, the algorithm proceeds to find similar states based on our similarity metric and similarity threshold. Weights are then defined which will determine the probability of retaining a state from the set of redundant states. These weights are determined by the probability of other non-similar states transitioning to the state of interest and then normalized. The state by which to retain is selected randomly in accordance to the probabilistic weights, and the rest of the similar states are erased from the state sequence. Vector $\tilde{\beta}$ is further updated by weakening the unselected similar states values in the vector.

Algorithm 1 Sample a State Sequence Containing No Redundant States

Initialize $\tilde{\beta} = \beta$, $(\tilde{x}_t)^{(m)} = (x_t)^{(m)}$, and define similarity threshold τ
Reorder $\{\theta_i : i \in (\tilde{x}_t)^{(m)}\}$ into new order $\{\theta_{I_i} : i \in (\tilde{x}_t)^{(m)}\}$ using random sampling without replacement where

- i corresponds to index the unique states existing in $(\tilde{x}_t)^{(m)}$
- I_i corresponds to the new index of state i in the new order $I = \{1, 2, 3, \dots\}$

while I is not an empty set **do**
  Let i correspond to the first I_i appearing in the new order I
  Calculate $\mathcal{D}\big(f(\theta_i), f(\theta_j)\big)$ for all $j \neq i$ where $j \in (\tilde{x}_t)^{(m)}$ ▷ Similarity metric.
  Define set $J = \{j : \mathcal{D}\big(f(\theta_i), f(\theta_j)\big) \leq \tau\}$ and set $J' = \{j : \mathcal{D}\big(f(\theta_i), f(\theta_j)\big) > \tau\}$
  for $j \in J$ **do**
    $\Pi_j = \sum_{i \in J'} \pi_{i,j}$ ▷ Weights depend on transition probabilities from non-similar states.
  end for
  Sample j^* from $\mathrm{P}(j^*)$ where $\mathrm{P}(j^* = j) = \Pi_j / (\sum_j \Pi_j)$ ▷ j^* is the redundant state to keep.
  Update $\tilde{\beta}_j = 0.1 * \beta_j$ for all $j \in J$ where $j \neq j^*$ ▷ Influence transition prior.
  Update $\tilde{x}_t = j^*$ for all $\{t : \tilde{x}_t \in J\}$ ▷ Influence data used for inference.
  Remove I_j from I for all $j \in (J \cup i)$. ▷ Prevent merging these states in future iterations.
end while
Output final $\tilde{\beta}$ and $(\tilde{x}_t)^{(m)}$ ▷ These will be used in next iteration of Gibbs sampling.

After implementing Algorithm 1, the sampling procedure is allowed to return to Step 1. Noticeably, every time this step is implemented, the algorithm begins with the originally sampled β and $(x_t)^{(m)}$, but ends with a $\tilde{\beta}$ and $(\tilde{x}_t)^{(m)}$ that encourages the transition matrix in Step 1 to promote transitions to non-redundant states and allow larger sample sizes for the available emission posteriors. Mathematically, the only adjustments made to Step 1 that reflect this dependency is

$$\begin{aligned}
\{\theta_i\}^{(m)} &\sim h_{\theta_i}(\theta_i | (\tilde{x}_t)^{(m-1)}, (y_t), H, G, \tilde{\beta}) \\
\{\omega_i\}^{(m)} &\sim h_{\omega_i}(\omega_i | (\tilde{x}_t)^{(m-1)}, (y_t), H, G, \tilde{\beta}) \\
\{\pi_i\}^{(m)} &\sim h_{\pi_i}(\pi_i | (\tilde{x}_t)^{(m-1)}, (y_t), H, G, \tilde{\beta}),
\end{aligned}$$

which preserves the Markov Chain structure of a Gibbs Sampler.

3.5 A Simulation Study

In this section, simulations are used to demonstrate the advantages of the proposed rHDP-HSMM method. The robustness and modeling accuracy is compared with the existing HDP-HSMM method. The simulation is designed as follows.

For each simulation, a sequence of observed data is generated with 30 total change points based on the distributions and parameters in Table 3.2. The emission parameters were specifically selected

as they feature some small overlap between their distributions.

	Emission		Duration	Transition		
Distribution	**Normal**		**Poisson**	**N/A**		
Parameter(s)	**Mean**	**Variance**	**Rate**	**State 1**	**State 2**	**State 3**
State 1	4	1	6	0	0.3	0.7
State 2	0	1	6	0.8	0	0.2
State 3	-4	1	6	0.4	0.6	0

Table 3.2: *The list of the true parameters in the hypothetical dataset with three different states.*

The generated sequence begins with a state being randomly selected from the three listed in Table 3.2. A length of duration is sampled from the selected state's duration distribution, which determines how many samples to draw from that state's emission distribution. Once the emission samples are collected, they are stored in the sequence, and the next state is sampled according the to that state's transition probability. The process is repeated 30 times to create a simulated sequence of "observed" data. An example of a simulated dataset can be observed in the Figure 3.3.

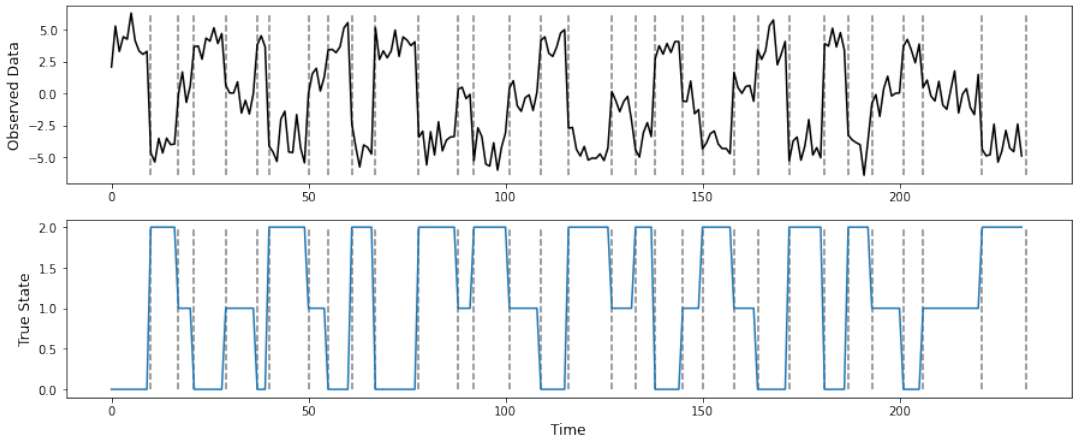
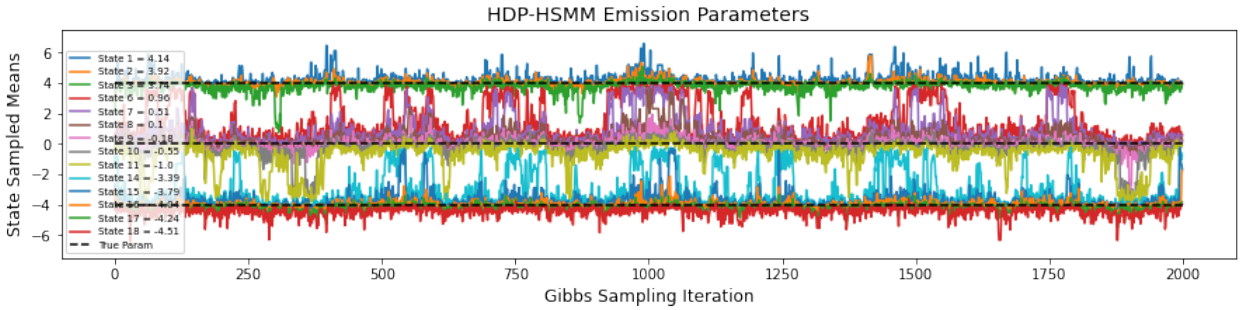


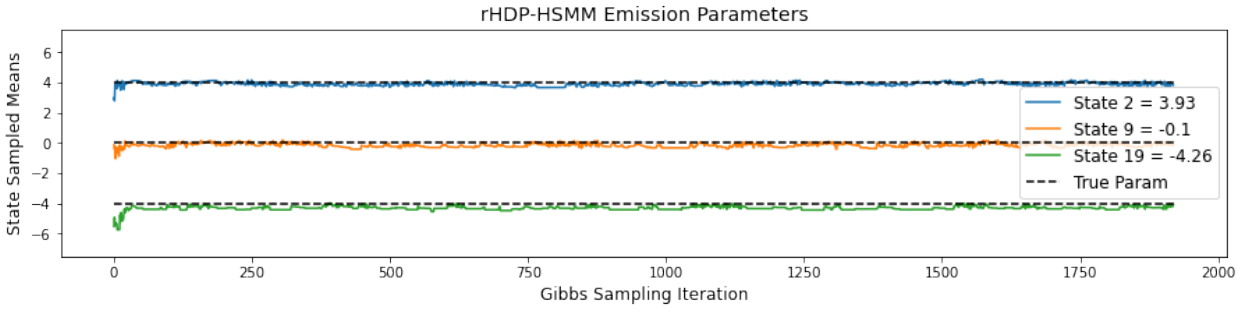
Figure 3.3: *Example of simulated data based on Table 3.2 and it's corresponding states.*

In each simulation, both the HDP-HSMM and the rHDP-HSMM are trained on the observed data with the same initial distributions and priors. The prior distributional forms were selected as to allow models to make use of conjugate relationships. Their parameters were selected as to ensure the true distributional parameters could be inferred with high probability. Each simulation's initial parameter values for the HDP-HSMM and rHDP-HSMM were drawn according to the selected prior. The maximum number of states for both models was set to 20. Each state's initial emission distribution was assumed Normal(μ, σ^2). The mean's prior distribution was set to $\mu \sim \text{Normal}(\mu_0 = 0, \sigma_0^2 = 4)$. The variance's prior distribution was set to $\sigma^2 \sim \text{InvGamma}(a_0 = 2, b_0 = 2)$. The initial duration distributions were assumed Poisson(λ), with prior $\lambda \sim \text{Gamma}(a_1 =$

$1, b_1 = 7)$. The transition distributions for each state is assumed to be $\pi_i \sim \text{Multinomial}(a_2)$, with the prior $a_2 \sim \text{Dirichlet}(a_3 = \mathbb{1}^{20})$. Both models had identifiability constraints implemented such as to order their states in increasing order of the posterior mean of their emission distribution. Furthermore, both models performed their respective Gibbs procedure over a maximum of 10000 iterations, or until their Gelman-Rubin statistic [17] reached less than 1.1. The burn-in period for both models was set to 100 iterations. Every 5th iteration of the sampled parameter chains was collected as to remove autocorrelation (resulting in a chain of 2000 length if convergence was not met). The rHDP-HSMM threshold for removing redundant states was set to 1.5. The posterior parameter values for each state was calculated as the mean of the most recent 20% of samples collected from the posterior parameter chains. The posterior sequence was selected to be the mode of the most recent 20% of samples collected from the posterior state sequence.



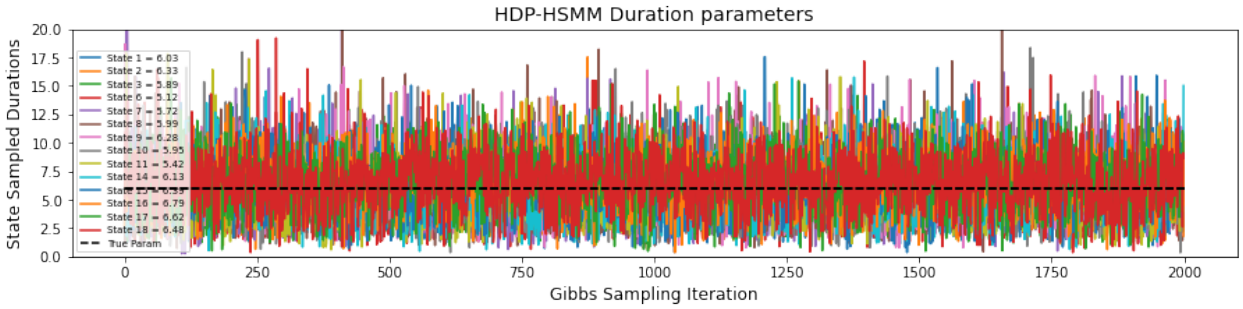
(a)



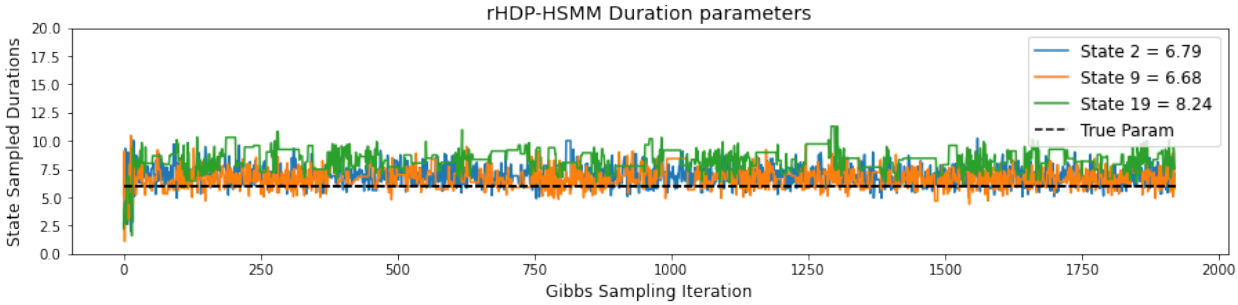
(b)

Figure 3.4: *HDP-HSMM versus rHDP-HSMM emission convergence on simulated data.*

The results of a single simulation are shown in Figures 3.4, 3.5, and 3.6. Figure 3.4 compares the HDP-HSMM and rHDP-HSMM's emission distribution convergence. The states shown in the plots are the states appearing in the final learned state sequence for each model. Each state is indicated by a different color. The true parameters are indicated by the dashed lines. While both models' posteriors are concentrated around the true parameters, the HDP-HSMM's posterior is multimodal for many states. Figure 3.4a shows how the many states rapidly switch which true



(a)

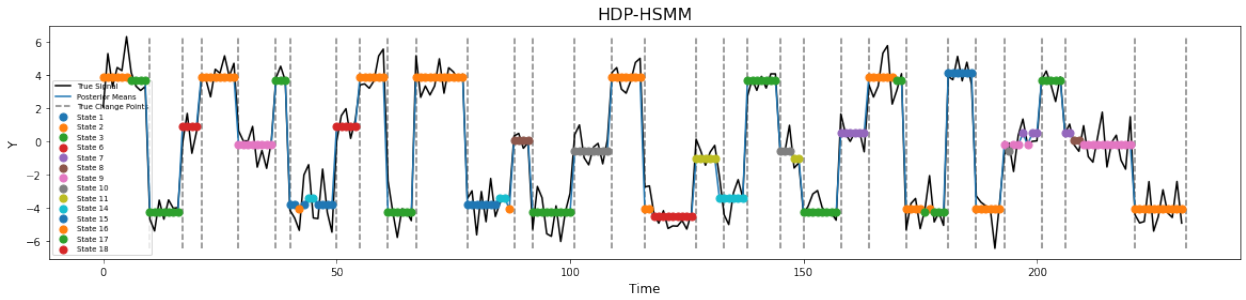


(b)

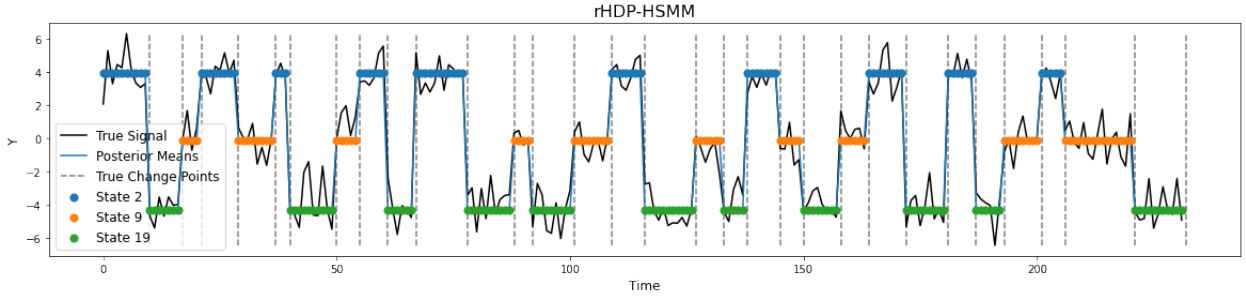
Figure 3.5: *HDP-HSMM versus rHDP-HSMM duration convergence on simulated data.*

state they want to encapture across sampling iterations. With regards to the duration, Figure 3.5 displays how both models posteriors are concentrated near the true duration. However, the variance of the HDP-HSMM's posterior samples is far larger than the variance of the rHDP-HSMM. This could be due to a large variation of samples being allocated to each state in the HDP-HSMM. To see this, Figure 3.6 shows the posterior state sequence estimated by both models. While the rHDP-HSMM distributes samples to each state under the constraint of removing redundant states, the HDP-HSMM's redundant states leave many states with very few samples left to estimate their duration. On a more positive note, both models are able to capture most of the true change points, however the HDP-HSMM leaves an impression of many more change point occurrences. Meanwhile, rHDP-HSMM clearly separates each of the 3 states from one another and captures only the locations of the true change points in the data.

The simulation is repeated 100 times, and the results are shared in Figure 3.7 and Table 3.3. Looking at the number of estimated states between the HDP-HSMM and the rHDP-HSMM, it is clear that the rHDP-HSMM's inference procedure removes states that would be otherwise present in a standard HDP-HSMM (Figure 3.7). In fact, 80 of the 100 simulations resulted in the rHDP-HSMM correctly inferring the true number of states. Furthermore, Table 3.3 shows that the rHDP-HSMM converged on average with fewer iterations than the HDP-HSMM. This table also shows that while



(a)



(b)

Figure 3.6: *HDP-HSMM versus rHDP-HSMM labeling of simulated data.*

both models are able to correctly capture all the true change points, the standard HDP-HSMM tends to estimate many more change points than the rHDP-HSMM. This is due to the redundancy issue, which the rHDP-HSMM eliminates through its modified inference procedure.

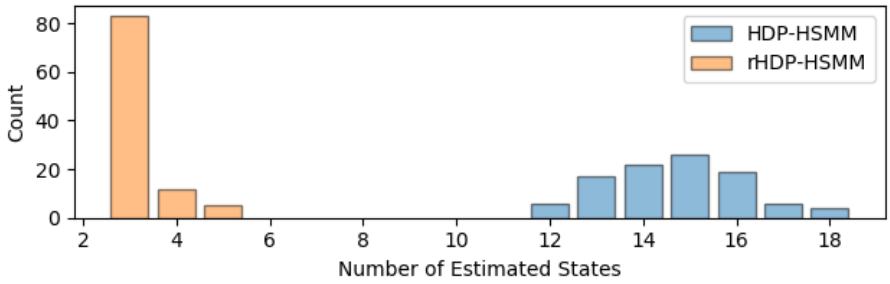


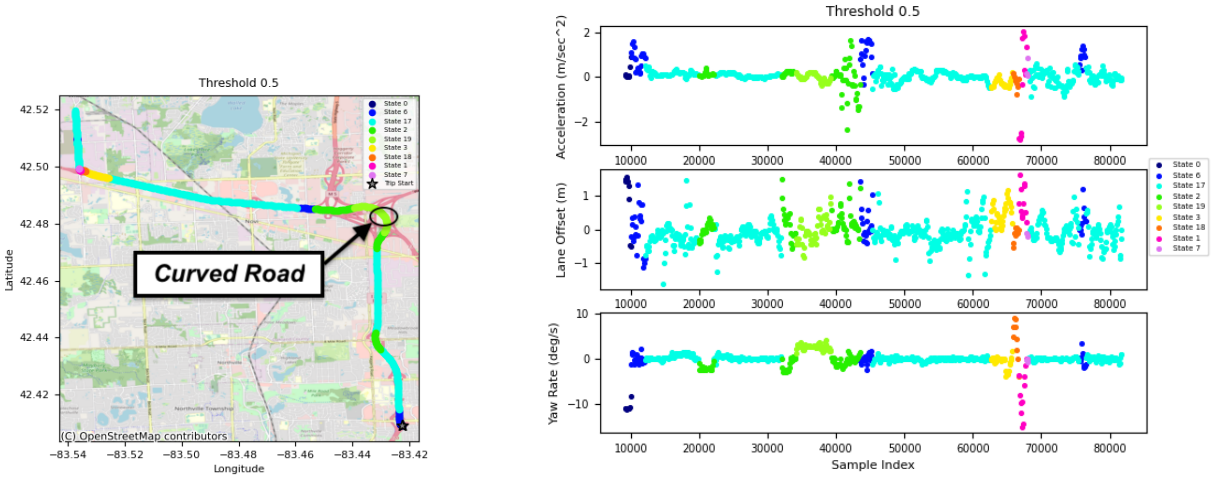
Figure 3.7: *The number of estimated states from 100 simulations comparing both a HDP-HSMM and the rHDP-HSMM.*

	HDP-HSMM	rHDP-HSMM
Num. of Converged Simulations	64	80
Avg. Num. of Gibbs Iterations	1372.5	993.5
Avg. Num. of Missed Change Points	1.0	1.3
Avg. Num. of Extra Change Points	15.0	1.5

Table 3.3: *The results of 100 simulations. Averaged values are calculated only from the iterations that converged.*

3.6 A Case Study on Naturalistic Driving Data

The benefit of the proposed rHDP-HSMM is demonstrated via the real-world application of modeling vehicle driving maneuver patterns. This type of modeling is useful for the development intelligent driving assistant systems and autonomous driving vehicles. The dataset analyzed in this study was collected by University of Michigan's Transportation Research Institute [39]. Several kinematic driving signals were collected from human-driven vehicles during their everyday activities. This naturalistic dataset is rich with information related to discover common driving maneuvers and behaviors. [69]. Signals are recorded on trip by trip basis, which begins when the vehicle is turned-on and ends when the vehicle is turned-off. An example of a trip can be seen in Figure 3.8.



(a) **(b)**

Figure 3.8: *A different segmentation of the road shown in Figure 3.1 labeled by a rHDP-HSMM under a threshold of 0.5.*

The kinematic signals of interest are acceleration, lane offset, and yaw rate. Acceleration and lane offset reflect a driver's intention of moving in the longitudinal and lateral directions

respectively. Yaw rate captures a driver's intention of of changing the forward direction of the car. Together, they form a multivariate time-series sampled at 10 Hz which should be highly correlated with human-driving behaviors. An example of the collected signals is shown in Figure 3.8. As maneuvers are expected to switch at a low frequency, the original data is down-sampled to 1 Hz by averaging every 10 data points.

Both the HDP-HSMM and a 0.5 threshold rHDP-HSMM are applied to trip shown in Figure 3.8 under the following setup. A 3-dimensional multivariate Gaussian distribution is used for the emission distribution ($Y \sim \text{MVN}(\mu, \Sigma)$). The priors to the emission mean and variance are selected as

$$\mu \sim \text{MVN}([0,0,0], [[1,0,0], [0,1,0], [0,0,1]])$$

$$\Sigma \sim \text{Inverse-Wishart}(2, [[1,0,0], [0,1,0], [0,0,1]]).$$

Each state's duration is assumed Poisson distributed ($D \sim \text{Poisson}(\lambda)$) with the prior $\lambda \sim \text{Gamma}(a = 1, b = 7)$. The identifiability constraints are constructed as to arrange the states in the order of smallest to largest mean and duration. The maximum number of states was limited to 20. The kinematic signals are normalized with respect to the signals observed during the trip. The learned emission means are transformed back to original space once the training is complete for analysis purposes.

The colors in Figure 3.8 represent the labeling results after training the 0.5 threshold rHDP-HSMM. Noticeably, the rHDP-HSMM segments the road into 9 states. Looking deeper at Figure 3.8b, it is clear that each state is primarily dictated by changes in yaw rate. Hence this model is able to capture portions of the road where various turning maneuvers are intended by the driver (Figure 3.8a). Comparing Figure 3.8a with the HDP-HSMM segmentation shown in Figure 3.1a, it is clear how the rHDP-HSMM merged the HDP-HSMM's 17 states into a more clear representation of maneuvers used on the road. An example of the rHDP-HSMM's estimated means versus the true recorded signals can be seen in Figure 3.9. In this graph, it can be seen the estimated states correspond mainly to the Yaw Rate signals, as the estimated mean values correspond almost one-to-one with the true vehicle signal. Considering the confidence intervals, the signal almost always lies within the interval.

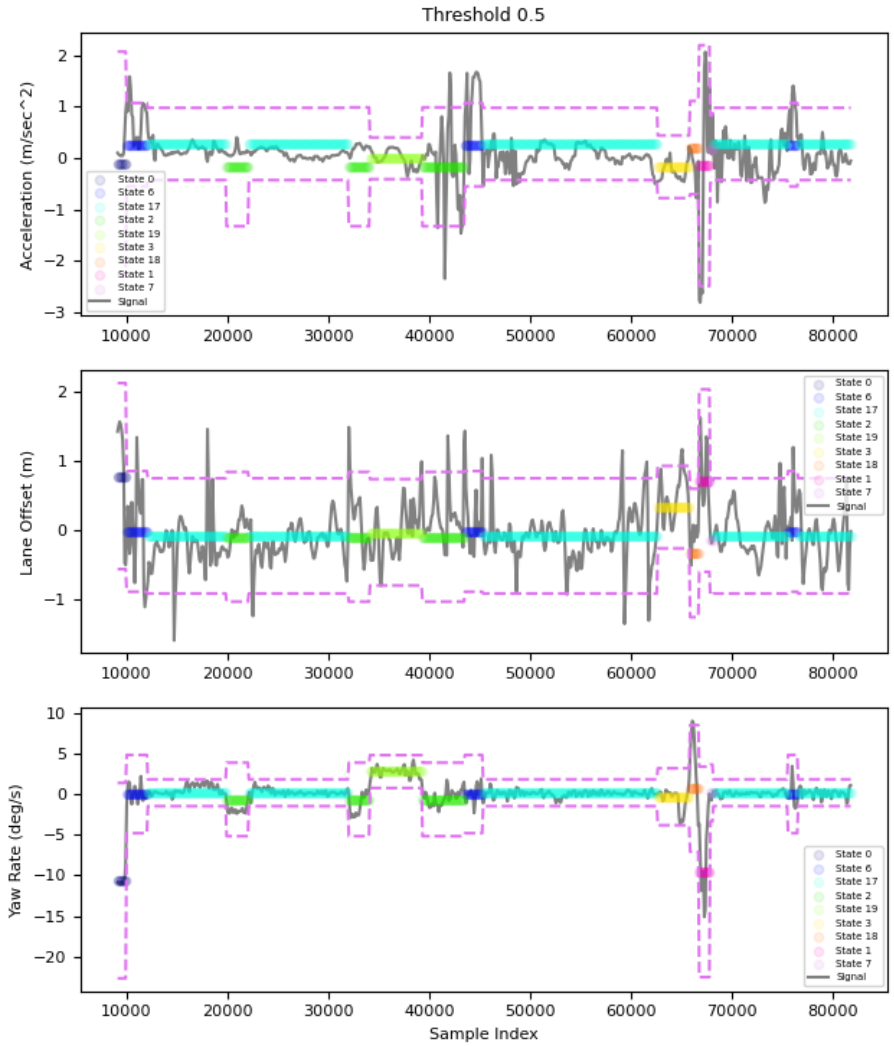
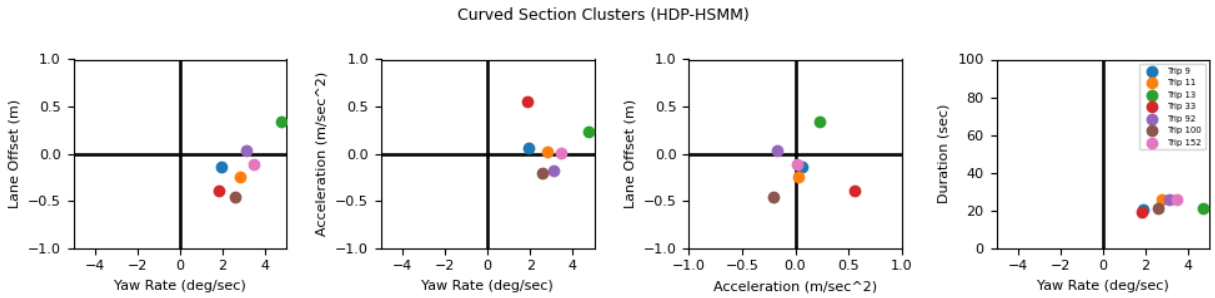
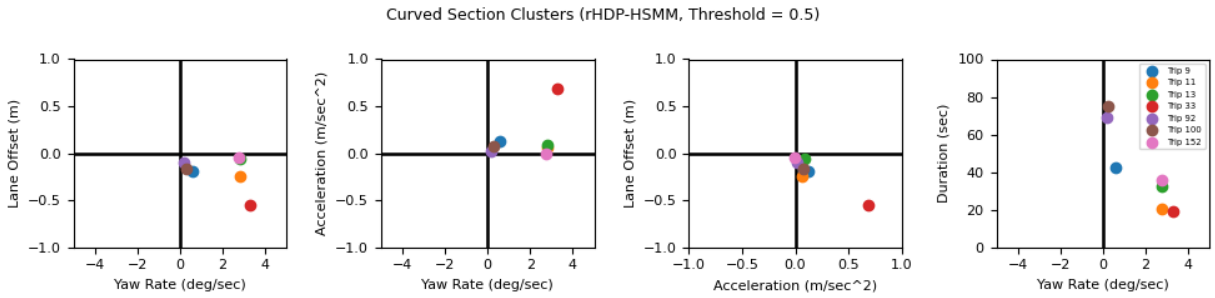


Figure 3.9: *The rHDP-HSMM estimated mean parameters plotted against the true recorded kinematic signals. The estimated means are the colored points on the graph, with each color corresponding to its respective state. The dashed line represents a two standard deviation interval around the estimated mean. The true signal is shown in gray.*

The rHDP-HSMM and HDP-HSMM are further compared in Figure 3.10 by using states obtained from the curved portion of the road marked in Figure 3.8a. Six other trips existed where the same driver drove on that part of the road. Hence, both the HDP-HSMM and the rHDP-HSMM are trained again on each of the other trips under the same initial parameters. The learned states from each model which occurred on the marked portion are analyzed in Figure 3.10. Figures 3.10a and 3.10b shows the emission means and durations learned by the HDP-HSMM and the rHDP-HSMM respectively. Interestingly, Figure 3.10b shows how the rHDP-HSMM concentrates the emission means in various quadrants of the graph. These quadrants relay a positive yaw rate, a negative lane offset, and a positive acceleration in all the learnt means. The concentration of these means in each



(a)



(b)

Figure 3.10: *Emission means corresponding to the kinematic signals from 7 different trips occurring on the curved portion of road shown in Figure 3.8. Figure 3.10a shows the means from the original HDP-HSMM, while Figure 3.10b shows the means from the proposed rHDP-HSMM.*

quadrant indicate a consistency in maneuvers among the various trips, which translates to a left turning action intended by the driver. This same conclusion is not easily recognizable in Figure 3.10a, as the HDP-HSMM loses this consistency in the learnt means. The difference in learning procedure between the HDP-HSMM and the rHDP-HSMM suggests that the HDP-HSMM's lack of concentrated means derives from the HDP-HSMM overestimating the number of states. As the rHDP-HSMM inference procedure merges similar states together, the emission means of each state can be inferred with a greater amount of data, providing both more consistent estimates and more consistent conclusions.

3.7 Discussion and Conclusion

The HDP-HSMM is a powerful model for discovering driving maneuver patterns from kinematic driving data. This paper details an extension to the HDP-HSMM in which this paper refers to as a robust HDP-HSMM (rHDP-HSMM). This model provides a solution to the inconsistency problem caused by the HDP prior. Looking through the lens of a weak-limit approximation of the

HDP prior, the problem typically occurs as the Dirichlet distribution takes no consideration for redundant states, which dilutes the underlying transition process. The rHDP-HSMM solves this issue by adjusting the sample from Dirichlet distribution by checking which states can be merged together. The model then scales down the weights which encourage transitions to redundant states. As a result, the rHDP-HSMM learns fewer redundant states and estimates longer state durations when compared to the original HDP-HSMM. This change leads to improved segmentation and more accurate transition probability representation, which is useful for the application of learning driving maneuvers.

Two case studies are presented to further demonstrate the ability of the proposed rHDP-HSMM over the HDP-HSMM. The first study is a simulation which utilizes 1-dimensional normal distributions for the emission function. The rHDP-HSMM demonstrates a clear improvement with regards to the posterior chains. The emission parameters converge much faster, the duration posteriors have far less variance than the HDP-HSMM's duration posterior, and finally the posterior state sequence presents far less change points than the HDP-HSMM's. Over the course of 100 simulations, the rHDP-HSMM out performs the HDP-HSMM in terms of convergence and having less extra change points relative to the truth.

The second study demonstrates of the effectiveness of the model in identifying and inferring driving maneuver patterns from a naturalistic dataset of kinematic signals. It is shown how the rHDP-HSMM's merging procedure reduces the number of states to describe a trip from 17 to 9 states when compared to a regular HDP-HSMM. The states are highly interpretable and now specifically capture portions of the road where various turning maneuvers are intended by the driver. In addition to this, the study also compares the results from multiple trips occurring on a curved portion of the road. The results show how the rHDP-HSMM consistently estimates similar emission distributions from multiple trips when compared to the original HDP-HSMM estimates.

In both studies, the rHDP-HSMM outperforms the HDP-HSMM in terms of estimation and consistency. This paper concludes that the rHDP-HSMM is worth applying to datasets where an HDP prior may be generating redundant states. Further inspection as to how to select the threshold may be required, however it is clear that the merging procedure within the model is still able to learn consistent and highly interpretable states for the study of driving maneuvers.

CHAPTER 4

Constrained Gaussian Processes for Airline Demand Prediction and Customer Behavior Inference

4.1 Introduction

In the field of revenue management, understanding consumer behavior is important for capacity control decisions. Capacity control decisions are made to optimally allocate the capacity of a resource to various consumers. For airline revenue management, airlines consider how much capacity should be allocated to various fare groups before deciding the price of each seat on a plane. This decision provides customers with various options to choose from when purchasing, which in turn impacts consumer behavior. Customers will respond differently depending on the options presented before them. The more relevant a product appears to a customer, the higher the probability of a purchase being made, i.e., if an option is within their preferred price range, a customer may make a purchase. Also, if the ticket price is slightly above their preferred range, there is still a chance for customers to make a purchase decision. It is always of interest for airlines to maximize their revenue. To take advantage of this probability, airlines can optimize their strategic decisions when determining the availability of various fares. The allocated availability of different fares will then impact the amount of purchases the airline might actually observe.

Capacity control decisions in traditional airline revenue management systems usually begin with a forecast of the customer demand for various fare groups, based on which, the number of seats to be sold to each fare group is determined through the airline's optimization process (i.e. how many seats should be allocated to each fare group). The initial demand forecasting step is of high importance with regards to the optimization [10]. In fact, the allocation of seats to each fare group sensitively affects historical customer purchasing data to be biased towards these allocation decisions. Hence, historical demand observations are often censored according to the airline's allocations to each group, and the demand estimation becomes difficult for unobserved scenarios falling outside of these historical allocation ranges.

4.2 Literature Review

In order to combat this censoring problem, demand has been modeled in a variety of ways. [19] classifies most methods into either single-class or multi-class methodologies. The single-class methodologies assume the demand in each fare group is independent of the demand in other fare groups. Hence the relationships among different fare groups are not utilized in these methods. Most of the single-class methodologies are well documented and surveyed in [65, 19, 52, 64, 68]. The notable methods include imputation techniques, Double Exponential Smoothing (DES), Expected Maximization (EM), and Projection Detruncation (PD). Among these, [68] showed the EM algorithm and PD methods to be more robust in estimating demand under censored scenarios, and the EM algorithm required less computational effort compared to the PD method. However, the EM and PD methods are still being challenged by ongoing research. [13] related the censoring problem to the field of survival analysis and derived a maximum likelihood method that takes into consideration the level of censorship in the data. The results showed that survival modeling was able to produce almost identical results to the EM algorithm, but provided more intuition behind the relationships between censorship and demand. [44] presented a Gaussian Process (GP) regression model to learn and extrapolate the demand. The results showed that GP regression to outperform DES, EM, and PD methods and was able to cope with important characteristics of realistic demand data, including nonlinear demand trends, variations in the total demand, lengthy periods of censorship, non-exponential inter-arrival times, and discontinuities/change-points in demand data.

The single-class methods assume customer behavior of one given fare group is independent of other fare groups being observed. In practice, this assumption over-simplifies the impact of airlines pricing decisions on consumer demand correlation among the different fare groups. For example, passengers will "buy down" if cheaper fares are available. [5] discusses observed demand being either *yieldable* demand or *priceable* demand. *Yieldable* demand is defined as the observed demand when the customer purchases the more expensive fare products, even when cheaper fares exist. *Priceable* demand is the demand observed from customers who are primarily concerned with the price, and will always purchase the lowest available fare. This perspective is interesting as it captures the characteristic of customers being either sensitive or insensitive to price. The difference is important as each type of customer demand should utilize different forecasting models.

This leads to a research interest in how to model customer demand correlations among different fare groups. One approach is the use of sell-up rate probabilities. These probabilities aim to capture the probability of customers 'sell-up' behavior, which describes the phenomena of a customer purchasing a fare outside of their preferred range. For example, [21] utilized fare ratios and observed demands to estimate sell-up rates. [14] provides an alternative sell-up rate

calculation using a multinomial logit model to consider cross-flight interactions. Sell-up rates aid the estimation of the dependent component of demands in that they can convert the demands from one fare group to another fare group [66]. However sell-up rates are generally difficult to obtain as sell-up occurrences are rarely captured during the booking process [6]. The data used to calculate market-specific sell-up rates is inherently sparse. System-wide aggregation methods are often used to make use of more data across the entire system. However estimating sell-up over an entire system results in estimates that are uncharacteristic of the market of interest. 'Buy-down' is a case where customers buy a cheaper fare class than they originally intended to buy. These probabilities are also of interest to calculate in practice for managing risk.

Many multi-class uncensoring methods exist that do not directly use the sell-up rates to model the demand correlations among fare groups. For example, [55] presented a multivariate linear mixed effects model for modeling random demands. The model captured both the time dimension and the pricing dimension of demand and related them through a shock variable. The shock variable induced correlation across both dimensions. This model estimates the parameters of demand via classical incomplete data framework and utilizes an EM algorithm to uncensor demand. [31] also represented the correlation between two different fare-classes via a covariance function. Their work extended the EM uncensoring algorithm to include correlation in the estimation.

This paper presents an uncensoring methodology which utilizes modern constrained GP regression [2] to estimate different types of demand for censored fare classes. The results of estimates are then converted to obtain both sell-up probabilities and buy-down probabilities. Hence, while these probabilities are generally difficult to obtain as these type of occurrences are rarely captured during the booking process, the results of this paper will show how a multi-fare constrained Gaussian Process model can be used to properly describe these probabilities. The contribution of this work will highlight a new perspective for more accurate estimation of various types of demand for multiple fare groups, and show how convert these demand estimates can be used to determine sell-up and buy-down probabilities.

The remaining sections will discuss the following. Section 4.3 describes the airline demand setting, and the difficulties involved in estimating demand. Section 4.4 describes the modeling framework, beginning with Gaussian Process Regression via a single-fare group setup, which is then improved by presenting a multiple-fare group setup. The constrained Gaussian Process Regression is presented in this section along with general constraints recommended for improving estimation. Section 4.5 describes the difficulties of estimating customer buy-down and sell-up behaviors, and shows how the described constrained Gaussian Process can be used to estimate these type of behaviors. A simulation example is provided in Section 4.6 to demonstrate how the constraints change the Gaussian Process predictions, and further highlight important parameters to the constraints. Section 4.7 demonstrates how the described constraints can be used to estimate

the demand from realistic airline data. Section 4.8 converts the airline data into the quantities needed to calculate the customer buy-down and sell-up probabilities, and demonstrates how the developed constrained Gaussian Process model can help infer these quantities. Finally conclusions are summarized in Section 4.9.

4.3 Data Structure

4.3.1 Data Structure

For the given flight route, there are multiple flights indexed by i ($i = 1, \ldots, n$), which contain sales from different fare groups indexed by j ($j = 1, \ldots, K$). The price of each group is denoted by p_j which is ordered $p_1 < p_2 < \cdots < p_K$; and their corresponding tickets' selling times are allocated as $\tau_{i,j}$. The amount of demand $y_{i,j}$ corresponding to each fare group is random, but strongly depends on the amount of selling time $\tau_{i,j}$. Figure 4.1 shows a simulated example of the described data. In order to maximize revenue, airlines are interested in optimizing the amount of selling time $\tau_{i,j}$. For this purpose, this requires an accurate estimation of demand for each fare group at various given selling times. This is difficult as some fare groups only experience certain types of time allocation $\tau_{i,j}$ to generate observable demand. For example, some fare groups j^* may have $\tau_{i,j^*} = 0$ as the fare group j^* was never allocated selling time for flight i. Thi translates to no available samples of demand y_{i,j^*}. In Figure 4.1, Groups 1 and 3 only have data where $\tau < 10$, implying that demand is unknown beyond that bound. For such fare groups j^*, it is desirable to infer its demand outside of its historical bounds by utilizing the information observed from other fare groups ($j \neq j^*$) that experienced different selling strategies.

In order to relay information from one group to another, certain assumptions are essential. For example, the demand decreases as price p_j increases, i.e., at the same selling time t, if groups j^- and j^+ relate with prices $p_{j^-} < p_{j^+}$, then the corresponding demands satisfy $E[y_{i,j^-}(t)] \geq E[y_{i,j^+}(t)]$. Furthermore, within a fare group, the corresponding demand often increases as time increases, i.e., for times $t^- < t^+$, the corresponding demand are assumed to have $E[y_{i,j}(t^-)] \leq E[y_{i,j}(t^+)]$. These relationships can also be observed in Figure 4.1. The next section will detail how Gaussian Process regression can incorporate these assumptions in modeling.

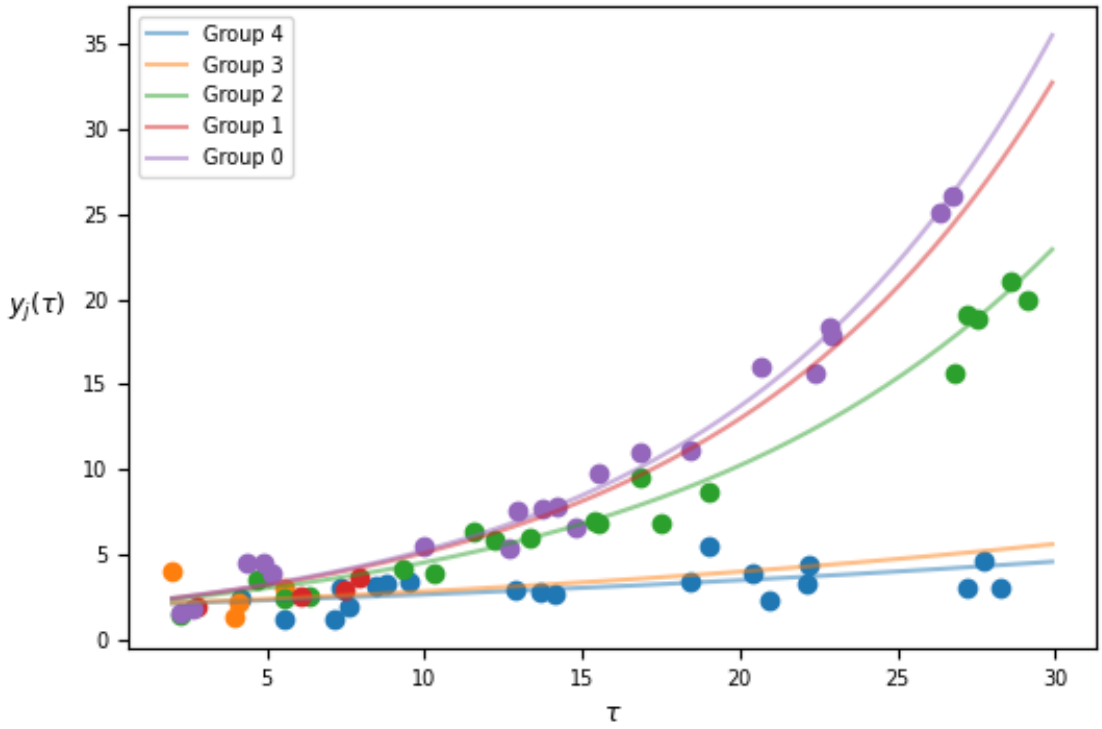


Figure 4.1: *Example of response data from different groups. Each group is represented by a different color. The true underlying function for each group is represented by the corresponding colored lines, while the observed data is represented by the points. Some groups have many observations at various values of τ, while other groups like Groups 1 and 3 only have observations at very limited values of τ*

4.4 Modeling Customer Demand

4.4.1 Single-Group: Gaussian Process Regression

Starting from a single group, the relationship between $y_{i,j}$ and $\tau_{i,j}$ is modeled as a Gaussian Process. To be consistent with standard GP notation, let $x_{i,j} = \tau_{i,j}$. All observations from a particular group j can be represented by column vector $\boldsymbol{y}_j \in \mathbb{R}^n$ corresponding to column vector $\boldsymbol{x} \in \mathbb{R}^n$, each of which denotes a vector with n observed $y_{i,j}$ and $x_{i,j}$ respectively. The distribution over vector $\boldsymbol{f}_j \in \mathbb{R}^n$ latent values corresponding to $\boldsymbol{x}$ is a multivariate Gaussian with

$$\boldsymbol{f}_j | \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}_j(\boldsymbol{x}), \boldsymbol{K}_j(\boldsymbol{x}, \boldsymbol{x}^\intercal)) \tag{4.1}$$

where $\boldsymbol{\mu}_j(\boldsymbol{x}) \in \mathbb{R}^n$ and $\boldsymbol{K}_j(\boldsymbol{x}, \boldsymbol{x}^\intercal) \in \mathbb{R}^{n \times n}$ denotes a Gram matrix of a selected kernel applied to the covariance matrices.

Given the observed data $(\boldsymbol{x}, \boldsymbol{y})$, one can estimate $\boldsymbol{y}_j^* \in \mathbb{R}^{n^*}$ corresponding to unobserved $\boldsymbol{x}^* \in \mathbb{R}^{n^*}$. Assume noise in the GP is represented by σ^2. Then the prediction distribution is still Gaussian with the mean and covariance as

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_j^* | \boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}_j] &= \boldsymbol{\mu}_j(\boldsymbol{x}^*) + \boldsymbol{K}_j(\boldsymbol{x}^*, \boldsymbol{x})[\boldsymbol{K}_j(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \boldsymbol{I}_n]^{-1}(\boldsymbol{y}_j - \boldsymbol{\mu}_j(\boldsymbol{x})) \\ \mathrm{Cov}[\boldsymbol{y}_j^* | \boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}_j] &= \boldsymbol{K}_j(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{K}_j(\boldsymbol{x}^*, \boldsymbol{x})[\boldsymbol{K}_j(\boldsymbol{x}, \boldsymbol{x}) + \sigma^2 \boldsymbol{I}_n]^{-1} \boldsymbol{K}_j(\boldsymbol{x}, \boldsymbol{x}^*) \end{aligned} \tag{4.2}$$

where $\boldsymbol{y}_j^* | \boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}_j$ is the predictive posterior given data $(\boldsymbol{x}, \boldsymbol{y}_j)$ and $\boldsymbol{I}_n$ is an identity matrix with n rows. Note that $\mathbb{E}[\boldsymbol{y}_j^* | \boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}_j] \in \mathbb{R}^{n^*}$ and $\mathrm{Cov}[\boldsymbol{y}_j^* | \boldsymbol{x}^*, \boldsymbol{x}, \boldsymbol{y}_j] \in \mathbb{R}^{n^* \times n^*}$

For the purpose of modeling airline demand, utilizing GP regression is advantageous because it can capture non-linear relationships of demand over time. Furthermore, the GP's predictive covariance matrix provides uncertainty for risk evaluation. [44] also presented a GP regression method to predict demand, which was able to outperform popular traditional estimation methods like Expected Maximization and Projection Detruncation. Different from the single-group GP regression model, we will form a constrained Gaussian Process (CGP) model to consider the relationships between different fare groups to facilitate prediction for groups with limited data.

4.4.2 Multiple-Groups: Notation

The single group setup of Gaussian Process regression limits the learning capabilities of the GP. Specifically, by not considering all groups simultaneously, there is no way to communicate information between different groups. To remedy this, this section will provide the setup and notation which will allow the modeling of all groups simultaneously.

Each flight $i = 1, \ldots, n$ provides demand samples for all fare groups $j = 1, \ldots, K$, even if the demand observation is 0 due to τ equating to 0. Hence, we can represent all demand samples into a single vector

$$\boldsymbol{y} = \left[y_{11}, y_{12}, \ldots, y_{1K}, y_{21}, \ldots, y_{2K}, \ldots, y_{n1}, \ldots y_{nK} \right]^\intercal \tag{4.3}$$

where $\boldsymbol{y} \in \mathbb{R}^{nK}$. The predictors corresponding to these observations can now be defined via the matrix

$$\boldsymbol{X} = \left[\widetilde{\boldsymbol{x}}_{11}^\intercal, \widetilde{\boldsymbol{x}}_{12}^\intercal, \ldots, \widetilde{\boldsymbol{x}}_{1K}^\intercal, \widetilde{\boldsymbol{x}}_{21\intercal}, \ldots, \widetilde{\boldsymbol{x}}_{2K}^\intercal, \ldots, \widetilde{\boldsymbol{x}}_{n1}^\intercal, \ldots \widetilde{\boldsymbol{x}}_{nK}^\intercal \right]^\intercal . \tag{4.4}$$

where $\boldsymbol{X} \in \mathbb{R}^{nK \times K}$ and each row of the matrix is defined by

$$\widetilde{\boldsymbol{x}}_{i,j} = \tau_{i,j} * \mathbf{e}_j^\intercal, \tag{4.5}$$

where $\mathbf{e}_j \in \mathbb{R}^K$ denotes a column vector with the j-th element value being 1 and the rest being 0. With this structure, $\widetilde{\boldsymbol{x}}_{i,j} \in \mathbb{R}^{1 \times K}$ is a row vector with the j-th column being the $\tau_{i,j}$ allocated to the

j-th group from observation i. The rest non-j columns are useful in that now the GP has a method by which to allow communication between the different groups.

4.4.3 Airline Demand Constraints

With the data set up like this, the constraints mentioned in Section 4.3 can now be mathematically described. First, we define a meaningful $\boldsymbol{X}$ (which we denote as $\boldsymbol{X}^m$) for which demand constraints must hold. To understand why $\boldsymbol{X}^m$ is meaningful, assume

$$\boldsymbol{X}^m = \tau * \boldsymbol{I}_K \tag{4.6}$$

where τ is some predefined time value and $\boldsymbol{I}_K$ is a $\mathbb{R}^{K\times K}$ identity matrix. In this case, the corresponding rows of the response vector $\boldsymbol{y}^m \in \mathbb{R}^K$ corresponds to the demand of each fare group j having time allocation τ. As all fare groups in $\boldsymbol{y}^m$ share the same level of allocated time, relationships between fare groups can be mathematically defined around $\boldsymbol{y}^m$. It should be noted that data need not be available at these locations, but rather the constraints are assumed to hold true at these locations.

With meaningful $\boldsymbol{X}^m$ and $\boldsymbol{y}^m$ defined, the demand constraints mentioned can be described via a linear operator $\mathcal{L}$ and bounds $a(\boldsymbol{X}^m)$ and $b(\boldsymbol{X}^m)$. The 3 constraints are as follows:

---

1. **Demand Positivity Constraint**: This constraint communicates the assumption that $0 \leq y_j$ for all fare groups j. Let $\boldsymbol{X}_1^m$ be defined by Equation 4.6, the bounds and linear operator that convey these constraints are

$$a_1(\boldsymbol{X}_1^m) = \boldsymbol{0}_K, \qquad b_1(\boldsymbol{X}_1^m) = \boldsymbol{\infty}_K, \qquad \mathcal{L}_1 = \boldsymbol{I}_K \tag{4.7}$$

   where $\boldsymbol{0}_K$ represents a vector of K zeroes and $\boldsymbol{\infty}_K$ represents a vector of K infinity values. Applying the bounds and linear operator to y_1^m corresponding to $\boldsymbol{X}_1^m$ results in the constraint $\boldsymbol{0}_K \leq \mathcal{L}_1 y_1^m$.

2. **Constraint Between Fare Groups**: This constraint relays the assumption that demand decreases as price increases. Let $\boldsymbol{X}_2^m = \tau * \boldsymbol{I}_{K-1}$. The constraint uses the following bounds and linear operator $\mathcal{L}_2 \in \mathbb{R}^{(K-1)\times K}$ on $\boldsymbol{y}_2^m$

$$a_2(\boldsymbol{X}_2^m) = \boldsymbol{0}_{K-1}, \qquad b_2(\boldsymbol{X}_2^m) = \boldsymbol{\infty}_{K-1}, \qquad \{\mathcal{L}_2\}_{jq} = \begin{cases} 1 & q = j \\ -1 & q = j + p \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

where j is the row of the matrix $\mathcal{L}_2$, q is the column of the matrix $\mathcal{L}_2$, and $p \in [1, \ldots, K-j]$ is a predefined step size. Each element of $\mathcal{L}_2$ is conveyed through $\{\mathcal{L}_2\}_{jq}$. The step size p allows for comparison between different fare groups. For example, if $p = 1$, then

$$\mathcal{L}_2 = \begin{bmatrix} 1 & -1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & -1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & 1 & -1 \end{bmatrix},$$

and each row j of $\mathcal{L}_2 \boldsymbol{y}_2^m$ represents the difference $y_j - y_{j+1}$. Considering the bounds, each row of $\boldsymbol{0}_{K-1} \leq \mathcal{L}_2 \boldsymbol{y}_2^m$ communicates the relationship $y_{j+1} \leq y_j$ for all $j = 1, \ldots, K-1$.

3. **Constraint Within a Fare Group**: This constraint is intended to communicate the assumption that for a given fare group j, demand should increase as the time allocation τ increases. This constraint first requires a small adjustment to Equation 4.6 in that 2 time allocations values τ^- and τ^+ are required where $\tau^- < \tau^+$. Let

$$\boldsymbol{X}_3^m = \begin{bmatrix} \tau^- * \boldsymbol{I}_K \\ \tau^+ * \boldsymbol{I}_K \end{bmatrix}. \tag{4.9}$$

where $\boldsymbol{X}_3^m \in \mathbb{R}^{2K \times K}$. With this setup, the corresponding $\boldsymbol{y}_3^m$ has rows $j = 1, \ldots, K$ having time allocation τ^-, and rows $j = K+1, \ldots, 2K$ having time allocation τ^+. With $\boldsymbol{X}_3^m$ defined, a linear operator can now be described to communicate the assumption that demand should increase as the time allocation τ increases. Define the following bounds and linear operator $\mathcal{L}_3 \in \mathbb{R}^{K \times 2K}$ on $\boldsymbol{y}_3^m$

$$a_3(\boldsymbol{X}_3^m) = \boldsymbol{0}_K, \qquad b_3(\boldsymbol{X}_3^m) = \boldsymbol{\infty}_K, \qquad \{\mathcal{L}_3\}_{jq} = \begin{cases} -1 & q = j \\ 1 & q = j + K \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

where $\{\mathcal{L}_3\}_{jq}$ represents each element of $\mathcal{L}_3$. This results in the linear operator having the form

$$\mathcal{L}_3 = \begin{bmatrix} -1 & 0 & 0 & 0 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & -1 & 0 & 0 & \ldots & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & 0 & -1 & 0 & \ldots & 0 & 0 & 1 \end{bmatrix},$$

with each row j of $\mathcal{L}_3 \boldsymbol{y}^m$ representing the difference $y_j(\tau^+) - y_j(\tau^-)$. The bounds and linear

operator relay the relationship $y_j(\tau^-) \leq y_j(\tau^+)$ for all $j = 1, \ldots, K$.

Each of the constraints above can mathematically describe commonly used airline demand assumptions. These definitions are useful as they can be used to understand and/or manipulate how different fare groups relate to one another. For example, while the bounds shown in this section are generally the either a zero vector or an infinity vector, we can modify the lower bound of Constraint 2 according to the economics concept of price elasticity.

4.4.3.1 Modifying Constraint 2 to Consider Price Elasticity

The price elasticity of demand is an important concept in the field of economics that measures the responsiveness of the quantity demanded of a product with respect to a change in its price. The formula for price elasticity of demand is typically represented as:

$$E = \frac{\%\Delta y}{\%\Delta p} = \frac{(y_{j_2} - y_{j_1})/y_{j_1}}{(p_{j_2} - p_{j_1})/p_{j_1}} \tag{4.11}$$

where:

- E represents the price elasticity of demand
- $\%\Delta y$ denotes the percentage change in quantity demanded between products with demand y_{j_1} and y_{j_2} respectively
- $\%\Delta p$ denotes the percentage change in price between products with prices p_{j_1} and p_{j_2} respectively.

For the purposes of this constraint, price elasticity can be used to improve estimation of the lower bound. Airlines typically know both the price is assigned to each fare group, and have an estimate of the price elasticity E. Typically, negative elasticity will be observed as it implies demand will increase as price decreases. If the elasticity and the price of each group is known, Equation 4.11 can be rearranged to highlight the following:

$$(y_{j_2} - y_{j_1}) \propto \frac{(p_{j_2} - p_{j_1})}{p_{j_1}} \times E \tag{4.12}$$

where the left hand side of the equation corresponds to the demand difference between two fare groups. This is analogous to the lower bound of Constraint 2 in Section 4.4.3 where each row j of $\mathcal{L}_2 \boldsymbol{y}_2^m$ represents the difference $y_j - y_{j+1}$. Thus the right hand side of Equation 4.12 can be used to

also define a clearer lower bound to each row j in Constraint 2 that was described in Equation 4.8:

$$[a_2(\boldsymbol{X}_2^m)]_j = \frac{(p_q - p_j)}{p_j} \times E. \tag{4.13}$$

4.4.4 Multiple-Groups: Constrained GP Regression

Here we borrow from [2] to combine the notation and constraints described in Sections 4.4.2 and 4.4.3, as they define method by which to constrain a GP with constraints represented by linear operators and bounds. Let $\mathcal{L}$ be a linear operator on realizations of $\boldsymbol{f} \sim \mathcal{GP}(\mu(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X})$, and $\mathcal{L}\boldsymbol{f}$ is still a GP [47, 42]. For two functions $a(\boldsymbol{X})$ and $b(\boldsymbol{X})$, where $a(\boldsymbol{X}) \leq b(\boldsymbol{X})$ for all $\boldsymbol{X}$, the constrained posterior GP on the event that $a(\boldsymbol{X}) \leq \mathcal{L}\boldsymbol{f} \leq b(\boldsymbol{X})$ can be reached approximately by relaxing the constraint to only hold at a finite set of inputs defined by $\boldsymbol{X}^m$ where $\boldsymbol{X}^m \in \mathbb{R}^{n_m \times K}$. Let $C(\boldsymbol{X}^m)$ denote the event that the constraint $a(\boldsymbol{X}^m) \leq \mathcal{L}\boldsymbol{y} + \varepsilon^m \leq b(\boldsymbol{X}^m)$ where $\varepsilon^m \sim \mathcal{N}(\boldsymbol{0}_{n_m}, \sigma_m^2 \boldsymbol{I}_{n_m})$ represents additive white noise at the finite set of inputs. Then the posterior predictive distribution of the constrained GP can be estimated using a lemma provided by [2]:

$$\boldsymbol{y}^* | \boldsymbol{X}^*, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{X}^m, C(\boldsymbol{X}^m) \sim \mathcal{N}(\boldsymbol{\mu}^* + \boldsymbol{U}(\boldsymbol{C} - \mathcal{L}\boldsymbol{\mu}^m) + \boldsymbol{V}(\boldsymbol{y} - \boldsymbol{\mu}), \boldsymbol{\Sigma}) \tag{4.14}$$

$$\boldsymbol{C} \sim \mathcal{TN}(\mathcal{L}\boldsymbol{\mu}^m + \boldsymbol{U}_1(\boldsymbol{y} - \boldsymbol{\mu}), \boldsymbol{V}_1, a(\boldsymbol{X}^m), b(\boldsymbol{X}^m)) \tag{4.15}$$

where $\mathcal{TN}(\cdot, \cdot, a, b)$ is a Gaussian $\mathcal{N}(\cdot, \cdot)$ conditioned on hyper-rectangle $[a_1, b_1] \times \cdots \times [a_{n_m}, b_{n_m}]$ and

$$\begin{aligned} \boldsymbol{U}_1 &= (\mathcal{L}\boldsymbol{K}_{\boldsymbol{x}^m,\boldsymbol{x}})(\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 \boldsymbol{I}_N)^{-1}, & \boldsymbol{V}_1 &= \mathcal{L}\boldsymbol{K}_{\boldsymbol{x}^m,\boldsymbol{x}^m}\mathcal{L}^\intercal + \sigma_m^2 \boldsymbol{I}_{n_m} - \boldsymbol{U}_1 \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}^m}\mathcal{L}^\intercal \\ \boldsymbol{U}_2 &= \boldsymbol{K}_{\boldsymbol{x}^*,\boldsymbol{x}}(\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}} + \sigma^2 \boldsymbol{I}_N)^{-1}, & \boldsymbol{V}_2 &= \boldsymbol{K}_{\boldsymbol{x}^*,\boldsymbol{x}^*} - \boldsymbol{U}_2 \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}^*} \\ & & \boldsymbol{V}_3 &= \boldsymbol{K}_{\boldsymbol{x}^*,\boldsymbol{x}^m}\mathcal{L}^\intercal - \boldsymbol{U}_2 \boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}^m}\mathcal{L}^\intercal \end{aligned}$$

$$\boldsymbol{U} = \boldsymbol{V}_3 \boldsymbol{V}_1^{-1}, \quad \boldsymbol{V} = \boldsymbol{U}_2 - \boldsymbol{U}\boldsymbol{U}_1, \quad \boldsymbol{\Sigma} = \boldsymbol{V}_2 - \boldsymbol{U}\boldsymbol{V}_3^\intercal$$

where the following shortened notation is used: $\boldsymbol{\mu}^* = \mu(\boldsymbol{X}^*)$, $\boldsymbol{\mu} = \mu(\boldsymbol{X})$, $\boldsymbol{\mu}^m = \mu(\boldsymbol{X}^m)$, and $\boldsymbol{K}_{\boldsymbol{x},\boldsymbol{x}^*} = K(\boldsymbol{X}, \boldsymbol{X}^*)$. Whenever the the linear operator is used on a vector, it is interpreted as elementwise. Further implementation calculations based on the lower triangular Cholesky factor are also provided in [2].

This lemma is powerful in that the Gaussian Process can now estimate under the constraints listed in Section 4.4.3. This results in a constrained Gaussian Process (CGP), that acknowledges how demand should behave relative to time and between different fare groups. For the specific scenario of missing data, these constraints can help guide the prediction to follow assumed relationships between different groups.

4.5 Modeling Customer Buy-Down and Sell-Up

4.5.1 Customer Buy-Down and Sell-Up Behaviors

Typically, the time allocation between fare groups overlap in a hierarchical manner, such that a higher fare group is always allowed to be sold at the times allocated to the lower fare groups. This implies $\tau_j \leq \tau_{j+1}$ for $j = 1, \ldots, K-1$. The hierarchical structuring of fare group availability allows for buy-down and sell-up information to be made available. To understand this, first let $t_0, t_1, \ldots, t_{M-1}$ represent M change points that indicate a change/shift in selling pattern among any of the K fare groups. Figure 4.2 displays a simulated example of 4 fare groups, where the buying rate shifts according to the lowest available price. The change points imply M non-overlapping segments, where the same fare groups are kept available within each segment. During each segment m, each fare group j takes on one of three states, which is represented by

$$Z_j(m) = \begin{cases} \text{E}, & \text{if group } j \text{ is being sold and considered } \textit{expensive} \text{ during segment } m \\ \text{C}, & \text{if group } j \text{ is being sold and is the } \textit{cheapest} \text{ available group during segment } m \\ \text{X}, & \text{if group } j \text{ was } \textit{not sold} \text{ during segment } m. \end{cases} \tag{4.16}$$

The state $Z_j(m)$ represents a comparison between group j's price with the rest of the fare groups that are available of a given segment m. Assuming customer behavior depends solely on what fares are shown to the customer, state $Z_j(m)$ indicates which type of buying behavior will occur for fare group j during segment m. For example, if fare group j is the cheapest price available (i.e $Z_j(m) = C$), it will attract much more demand as opposed to if a cheaper fare group was available (in which case $Z_j(m) = E$). Hence state $Z_j(m)$ determines the rate at which demand will arrive.

State $Z_j(m)$ is also useful for understanding customer buy-down and sell-up behavior. Consider the quantity of customers buying tickets for fare group j when $Z_j(m) = C$ by y_j^{C}. Demand outflow can be described via the states. In the scenario that the airline lowered the price to a cheaper fare group, some portion of y_j^{C} may purchase down (buy-down) to the cheaper group, while the remaining portion may choose to stay in fare group j (due to the extra perks belonging to fare group j). The portion of demand that would stay in fare group j given a cheaper fare is captured when state $Z_j(m) = E$, and can thus be represented by y_j^{E}. Unfortunately, the remaining portion that would buy-down to a cheaper fare (call it y_j^{D} as they are *dependent* on the price) can never be directly observed by airlines, as no state $Z_j(m)$ is able to capture it. However, the dependent demand can be solved for based on its relationship to y_j^{C} and y_j^{E} which is described by the following equation:

$$y_j^{\text{C}}(\tau) = y_j^{\text{E}}(\tau) + y_j^{\text{D}}(\tau). \tag{4.17}$$

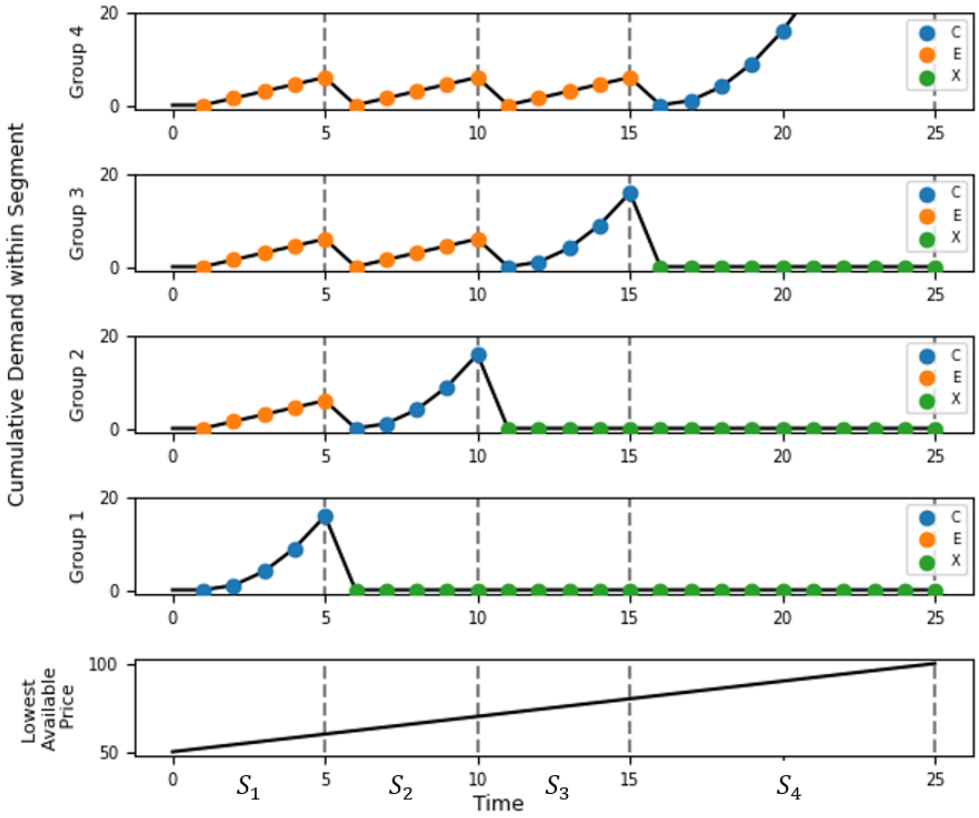


Figure 4.2: *An example displaying how the state of each fare group $Z_j(m)$ relates to the lowest available price for each segment m of time. Each segment m has S_m duration, and the value of $Z_j(m)$ for each fare group is labeled in the legend. For each segment, the arrival rate changes depending on the state of the fare group.*

Equation 4.17 represents demand outflow from group j and only holds true if all demand types have the same τ value. Solving for each of these terms is useful as they can be used to calculate the probability a customer in fare group j will buy-down to a cheaper group if it is also shown to the customer. This probability can be calculated as

$$\text{P(Group } j \text{ Buying Down } |\tau) = \frac{y_j^{\text{D}}(\tau)}{y_j^{\text{C}}(\tau)} = \frac{y_j^{\text{C}}(\tau) - y_j^{\text{E}}(\tau)}{y_j^{\text{C}}(\tau)}. \tag{4.18}$$

Another perspective of the quantity y_j^{C} is the demand inflow. That is some portion of its demand comes from the more expensive fare group $j+1$ via buy-down, while the remaining portion of its demand is new external demand that only purchases because they can afford the fare group j. The demand that bought down from the next highest fare group can be represented by y_{j+1}^{D}. However, by opening up the cheaper fare group j, new demand is attracted as more customers can now afford this product. The *new* customers that come because fare group j is now available can

be represented by y_j^{N}. By this logic, y_j^{C} both takes demand from the fare group above it and brings in new demand. This relationship can be described via the following equation:

$$y_j^{\mathrm{C}}(\tau) = y_{j+1}^{\mathrm{D}}(\tau) + y_j^{\mathrm{N}}(\tau). \tag{4.19}$$

Equation 4.19 represents demand inflow into group j and only holds true if all terms share the same τ value. Noticeably, the term $y_{j+1}^{\mathrm{D}}(\tau)$ in Equation 4.19 captures how much demand would actually sell-up to the next highest fare group $j+1$ if the airline were to make fare group j unavailable to customers. This is useful as the probability of a customer in fare group j selling-up to group $j+1$ can now be described as:

$$\mathrm{P}(\text{Group } j \text{ Selling Up } |\tau) = \frac{y_{j+1}^{\mathrm{D}}(\tau)}{y_j^{\mathrm{C}}(\tau)} = \frac{y_{j+1}^{\mathrm{C}}(\tau) - y_{j+1}^{\mathrm{E}}(\tau)}{y_j^{\mathrm{C}}(\tau)}. \tag{4.20}$$

where the numerator represented via Equation 4.17. This paper proposes predicting the values of $y_j^{\mathrm{C}}(\tau)$ and $y_j^{\mathrm{E}}(\tau)$ at some common τ value via the use of the presented constrained Gaussian Process model shown so far in order to estimate Equations 4.18 and 4.20.

4.5.2 Data Definitions for Modeling Buy-Down and Sell-Up

Section 4.4 demonstrated how to use CGPs for modeling airline demand without consideration of buy-down and sell-up. To employ the model to estimate under buy-down and sell-up assumptions, we propose representing the time allocations via their respective $Z_j(m)$ states. Denote each segment $m = 1, \ldots, M$ to have duration time

$$S_m = t_m - t_{m-1}, \tag{4.21}$$

where $t_0 = 0$, $t_M = T$, and $\sum_{m=1}^{M} S_m = T$. Assume the historical demand for each fare group j can be segmented additively according the each segment m and the state $Z_j(m)$:

$$y_j(T) = \sum_{m=1}^{M} y_j^{Z_j(m)}(S_m) \tag{4.22}$$

where $y_j^{Z_j(S_m)}(S_m)$ represents the observed demand according to state $Z_j(m)$ that occurred during segment m of group j.

Fare groups are only available for sell when $Z_j(m) \in \{\mathrm{E}, \mathrm{C}\}$. When $Z_j(m) = \mathrm{X}$, then $y_j^{Z_j(m)}(S_m) = 0$. Hence each fare group's time allocation can be represented by

$$\tau_j = \tau_j^{\mathrm{E}} + \tau_j^{\mathrm{C}},$$

where

$$\tau_j^{\mathrm{E}} = \sum_{m=1}^{M} S_m \times I(Z_j(m) \in \{\mathrm{E}\}), \tag{4.23}$$

$$\tau_j^{\mathrm{C}} = \sum_{m=1}^{M} S_m \times I(Z_j(m) \in \{\mathrm{C}\}), \tag{4.24}$$

and $I(\Omega)$ is an indicator function that equals 1 if event Ω is satisfied, and 0 otherwise.

Knowing the τ_j^{E} and τ_j^{C} partitions, the demand observations can also be decomposed into the two states:

$$y_j(\tau_j) = y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}}) + y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}}), \tag{4.25}$$

where $y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}})$ and $y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}})$ represent the observed demand under states E and C respectively for times τ_j^{E} and τ_j^{C}. As such, the CGP can be built around the considering the different components.

4.6 Numerical Study

To demonstrate the learning effects of transitioning from a multi-group setup of Gaussian Process modeling to constrained Gaussian Process modeling, an example is constructed such that a simulated response variable increases as the group number increases. Samples are generated from the following function for $2 \leq \tau \leq 30$ randomly:

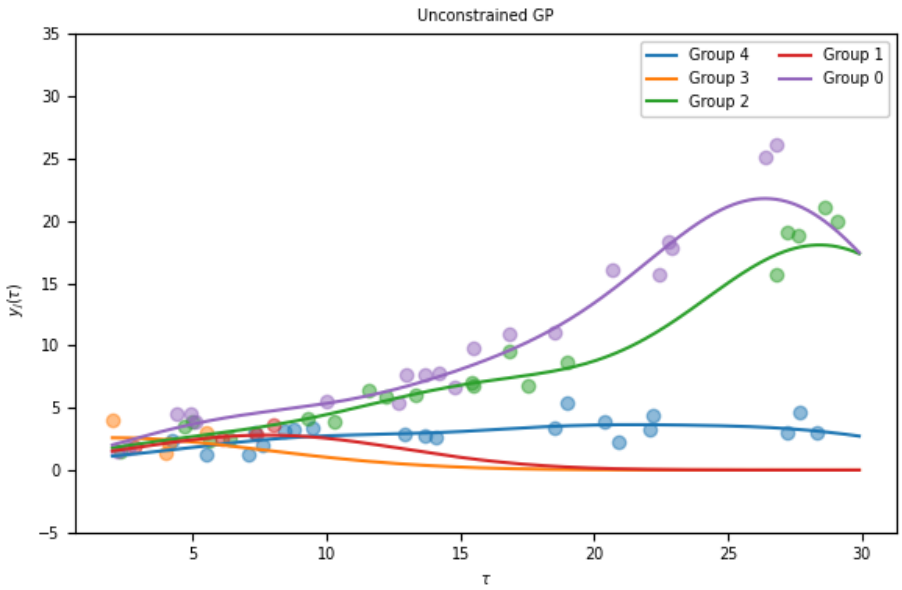
$$y_j^{\mathrm{Sim}}(\tau) = \beta(1 + r_j)^{\tau} + \varepsilon, \tag{4.26}$$

where $\beta = 2$ corresponds to the value of $y_j^{\mathrm{Sim}}(0)$, and r_j corresponds to the jth element of $[0.101, 0.098, 0.085, 0.035, 0.028]$, arranged such that smaller j will yield higher values of $y_j^{\mathrm{Sim}}(\tau)$. The variable $\varepsilon \sim \mathrm{N}(0, 1^2)$ represents random noise. If noise is not present, the function is simply an exponential growth curve, and smaller values of j receive larger response values. An example of the data from different groups can be seen in Figure 4.1.

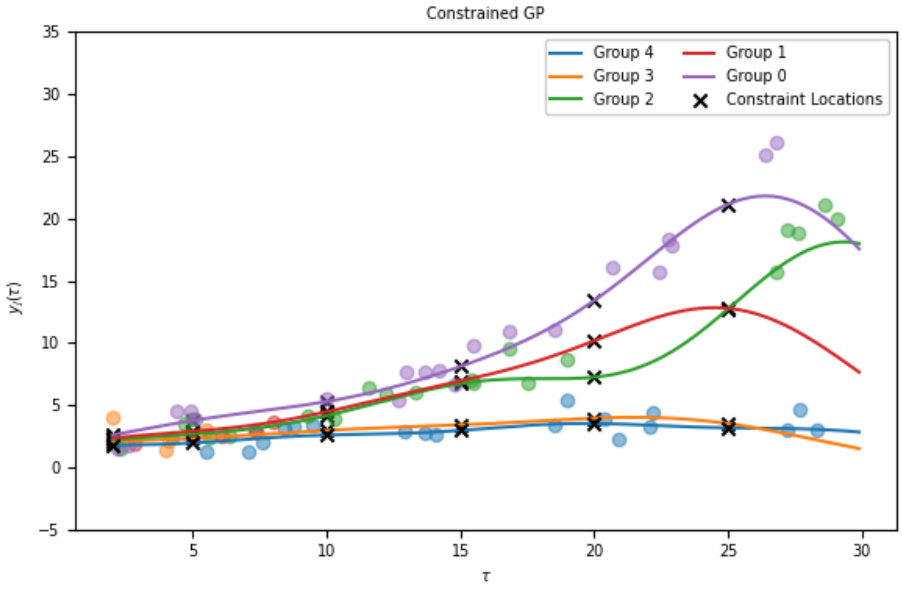
In Figure 4.1, Groups 1 and 3 only have data for $\tau \leq 10$. Hence training an unconstrained multi-group GP for each will result in extrapolation for these groups (Figure 4.3a). Extrapolating for values far beyond the observed data will result in estimates converging towards the prior (which is 0). The Gaussian Process that generated Figure 4.3a was constructed as follows. A radial basis covariance function is used

$$K(\boldsymbol{x}, \boldsymbol{x'}) = \sigma_k^2 * \exp(\frac{-||\boldsymbol{x} - \boldsymbol{x'}||^2}{2\ell^2}),$$

where variance is $\sigma_k^2 = 1$ and the length-scale parameter being $\ell = 5$. The prior mean is set to 0,



(a) *Fitted results of the Gaussian Process with no constraints.*



(b) *Fitted results of a constrained Gaussian Process.*

Figure 4.3: *Simulation results demonstrating the performance of a Gaussian Process with no constraints versus a constrained Gaussian Process utilizing only Constraint 2 described in Section 4.4.3.*

and the observation's noise is assumed to be $\sigma^2 = 1/3$. A GP for each group is trained under the same parameters using only the available data of the group currently being trained.

To provide more reasonable estimates for Groups 1 and 3, information from other groups can be

utilized and transferred via Constraint 2. Figure 4.3b demonstrates how implementing a CGP with only Constraint 2 described in Equation 4.8 greatly improves estimation for these groups. In this figure, it the constraints only hold at the defined $\boldsymbol{X}^m$ locations (denoted by the X-markers). Hence, when $\tau = 30$, the relationship between Group 1 and 2 inverts as Group 1 attempts tend back towards the prior. This highlights the importance of defining the meaningful $\boldsymbol{X}^m$ as your constraints will only truly hold at these locations.

The constrained Gaussian Process that generated Figure 4.3b was constructed to use the same parameters as the unconstrained GP. The only additional parameters are defined by the constraint inputs. The finite set of locations where the constraints are assumed to hold are set up in a manner similar to Equation 4.9 with $\tau \in [2, 5, 10, 15, 20, 25]$, resulting in

$$\boldsymbol{X}^m = \begin{bmatrix} 2 * \boldsymbol{I}_K \\ 5 * \boldsymbol{I}_K \\ 10 * \boldsymbol{I}_K \\ 15 * \boldsymbol{I}_K \\ 20 * \boldsymbol{I}_K \\ 25 * \boldsymbol{I}_K \end{bmatrix}.$$

The linear operator $\mathcal{L}_2$ defined in Section 4.4.3 is applied to each τ value and combined into a single linear operator as to apply the constraint to the different τ values. The bounds for the constraints between fare groups are selected to be $a_2(\boldsymbol{X}^m) = \boldsymbol{0}$ and $b_2(\boldsymbol{X}^m) = \boldsymbol{1}10000$ for each value of τ. The constraint's variance is selected to be $\sigma_m^2 = 0.001$. The constrained GP is trained using all available data from all groups simultaneously. This one constraint greatly aids the estimation of Groups 1 and 3 for $\tau > 10$ in Figure 4.3b in comparison to the regular GP in Figure 4.3a. To further improve estimation, adjustments of the lower bounds can also be made, which will be explored in the next section.

4.7 Case Study: Airline Demand Estimation

A case study is presented to highlight the capability of the methodology described in Section 4.4. In this case study, historical airline booking data from a particular customer partition is used train a CGP under this paper's constraints. The dataset contains of the amount of bookings purchased from different fare groups for various flights of the same route. The dataset also contains the respective amount of time allocated towards selling each fare group of each flight. The times are standardized to fall between $\tau_{i,j} = 0$ (no time allocated towards selling the fare group) and $\tau_{i,j} = 1$ (maximum amount of time possible spent towards selling the fare group). The fare groups j are numbered

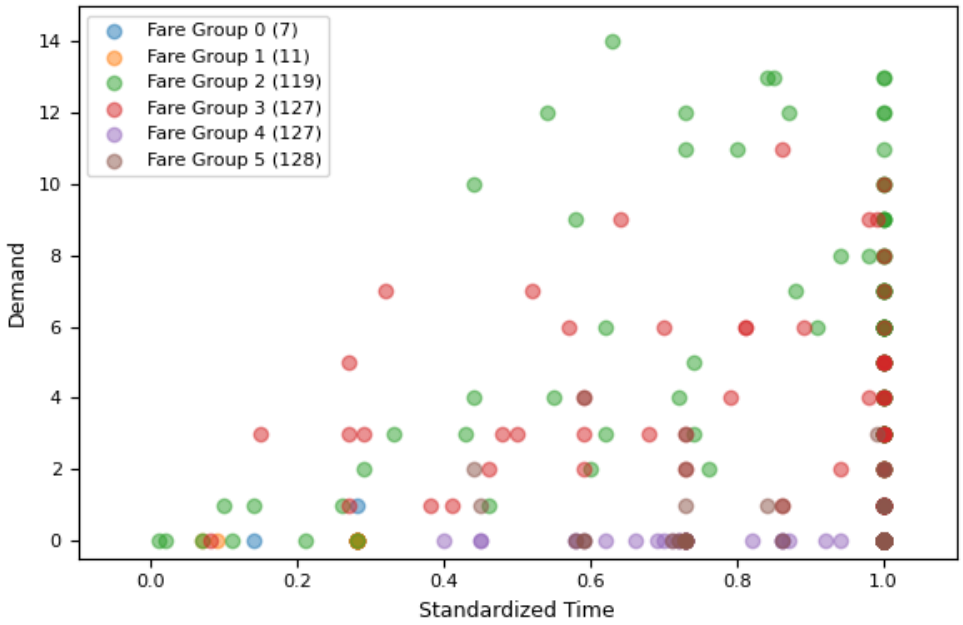
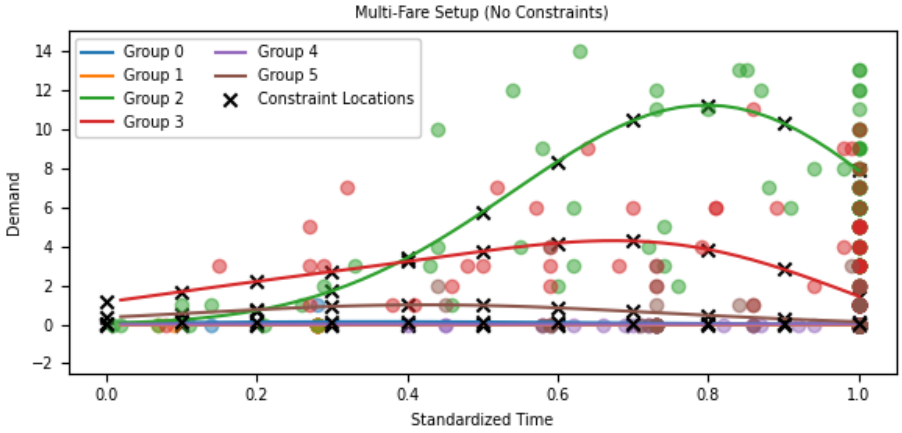


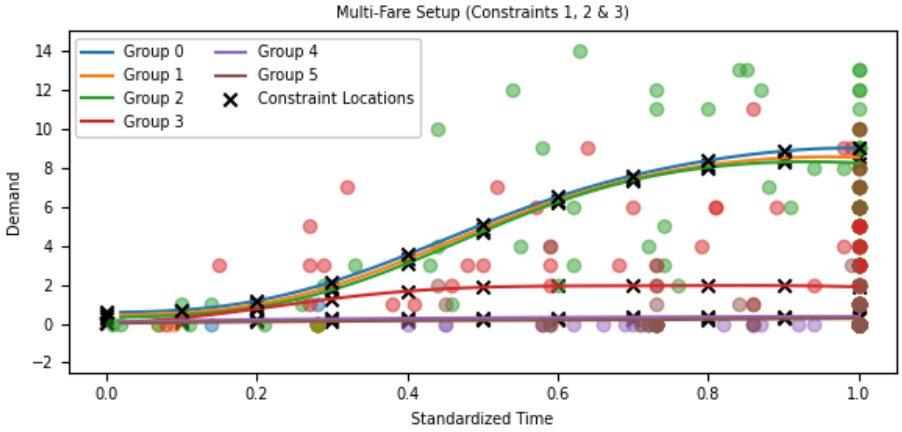
Figure 4.4: *Example of airline booking data from different fare groups. Each group is represented by a different color. The number of samples collected for each fare group is shown in the legend in parenthesis.*

such that smaller fare group ID's corresponds to cheaper prices.

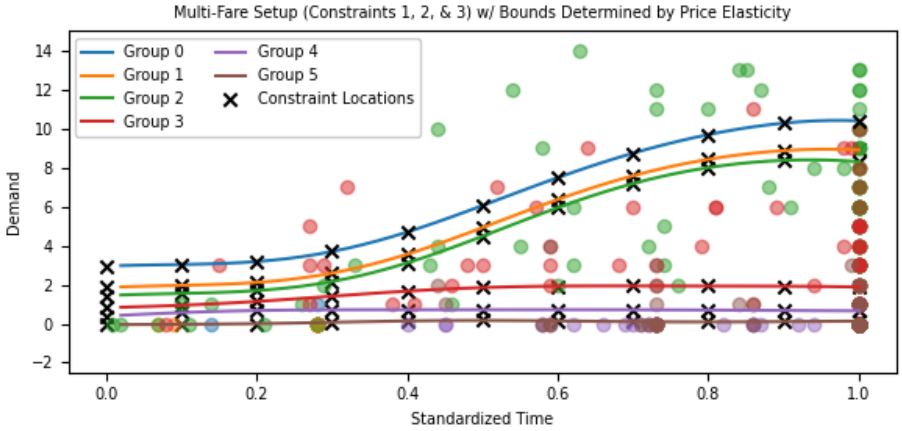
An example of the data can be observed in Figure 4.4, where the y-axis is the number of bookings observed for each fare group of interest, and the x-axis represents the standardized time allocated to each fare group. In this figure, certain fare groups historically have limited selling strategies in comparison to other fare groups. For example, Fare Groups 0 and 1 never experience time allocations greater than 0.3, while Fare Group 2 tends to get allocations ranging anywhere between 0 and 1. In order to make optimal decisions as to how much time should be allocated to each fare group, airlines need to accurately estimate the amount of demand each fare group can bring at any time allocation. As can be seen, Fare Groups 0 and 1 require much extrapolation with little-to-no information available for these groups at times greater than 0.3. To properly estimate demand for these groups, a CGP will trained utilizing the constraints described in this paper. The constraints will allow for information from Fare Groups 2 and 3 to help guide the estimation for groups with little to no information.



(a) *Multi-Fare GP setup with no constraints applied.*



(b) *Multi-Fare CGP setup with Constraints 1, 2, and 3 listed in Section 4.4.3.*



(c) *Multi-Fare CGP setup with Constraints 1, 2, and 3 listed in Section 4.4.3, but the constraint lower bounds are determined according to price elasticity.*

Figure 4.5: *The results from a multi-fare GP, CGP, and CGP with modified constraints trained on airline booking data.*

To show the impact of the constraints, first an unconstrained Gaussian Process is trained with a radial basis covariance function with variance $\sigma_k^2 = 1$ and the length-scale parameter being $\ell = 0.3$. The assumed observation's noise is $\sigma^2 = 1$. The prior mean is set to be 0. Figure 4.5a demonstrates the estimation curves from the GP using a multi-fare group setup. This figure shows that Fare Groups 0 and 1 having less demand estimated in comparison to Fare Groups 2 and 3 as their predictions are tend towards the prior mean. This is unrealistic as Fare Groups 0 and 1 are cheaper than Groups 2 and 3. Hence the predictions by this GP are nonsensical due to extrapolation with little-to-no data.

Applying constraints to the Gaussian Process can guide the predictions to be more realistic. To show this, a CGP is trained using the same parameters as the GP, however constraints 1, 2, and 3 from Section 4.4.3 are applied. The set of locations where the constraints are assumed to hold can be described to be 0.1 step intervals in a manner similar to Equation 4.9 with τ values equating to $0, 0.1, 0.2, \ldots, 0.9, 1$. However, as to ensure $\tau = 0$ has meaning, the non-diagonal terms of the $\boldsymbol{X}^m$ matrix are converted to -1, while keeping the diagonal terms providing information related ot τ. This is done via the following formulation:

$$\boldsymbol{X}^m = \begin{bmatrix} 0 * \boldsymbol{I}_K - 1 * (\boldsymbol{1}_{K,K} - \boldsymbol{I}_K) \\ 0.1 * \boldsymbol{I}_K - 1 * (\boldsymbol{1}_{K,K} - \boldsymbol{I}_K) \\ 0.2 * \boldsymbol{I}_K - 1 * (\boldsymbol{1}_{K,K} - \boldsymbol{I}_K) \\ \vdots \\ 1 * \boldsymbol{I}_K - 1 * (\boldsymbol{1}_{K,K} - \boldsymbol{I}_K) \end{bmatrix} \tag{4.27}$$

where $\boldsymbol{1}_{K,K}$ denotes a $\mathbb{R}^{K \times K}$ matrix of ones. The linear operators and bounds defined in Section 4.4.3 are combined into a single linear operator as to apply each type of constraint to each of the different τ values. The constraint's variance is selected to be $\sigma_m^2 = 0.001$ as to ensure the constraints are held tightly.

Figure 4.5b shows the results of the described constrained GP. The locations where the constraints are assumed to hold true can be observed by the x-markers, while each fare group's prediction is color coded. Under this setup, Fare Groups 2 and 3 are now positively monotonic with respect to time as opposed to using the regular GP. Furthermore, Fare Groups 0 and 1's predictions are now slightly greater than Fare Group 2's prediction, as the constraints let the CGP know these values should be higher than Fare Group 2. This results in more realistic demand estimates for all groups in general. While the regular GP has only the prior to infer from for groups with little-to-no data, the CGP utilizes the constraints to relate the observations from Fare Group 2 to aid the estimation of bookings in other groups.

The modified Constraint 2 described in Section 4.4.3.1 can be used to solve the issue of Figure

4.5b, in which the resulting GP estimates of Fare Groups 0 and 1 to be almost identical to Fare Group 2. Assume price elasticity $E = -1$ and assume the price of each fare group is known such that it increases from Fare Group 0 to Fare Group 5, then the adjustments to the CGP predictions after using the lower bounds described in Equation 4.13 can be observed in Figure 4.5c. In this figure, Fare Groups 0 and 1 are relatively more distinguishable from Fare Group 2 in accordance to the price differences between the groups.

The use of Gaussian Processes for demand estimation also allows for uncertainty to be be quantified. Equation 4.14 provides details for estimating both the predictive mean and the predictive covariance matrix as well. The diagonal terms of the covariance matrix can be used to describe uncertainty of the predictive mean. Figure 4.6 shows demand predictions from Fare Groups 0, 2, 3, and 5 and alongside a 2-standard deviation confidence interval around the mean. The intervals get wider as more uncertainty is present and narrower as the uncertainty drops. For example, Figure 4.6 shows how the intervals for Group 0 at $\tau > 0.3$ increases as little data is available in that region. The interval is not extremely wide however as the estimation is guided by the data available by Group 2. These confidence intervals are beneficial for airline optimization as uncertainty is valuable for optimization procedures deciding how much time should be allocated to each fare group.

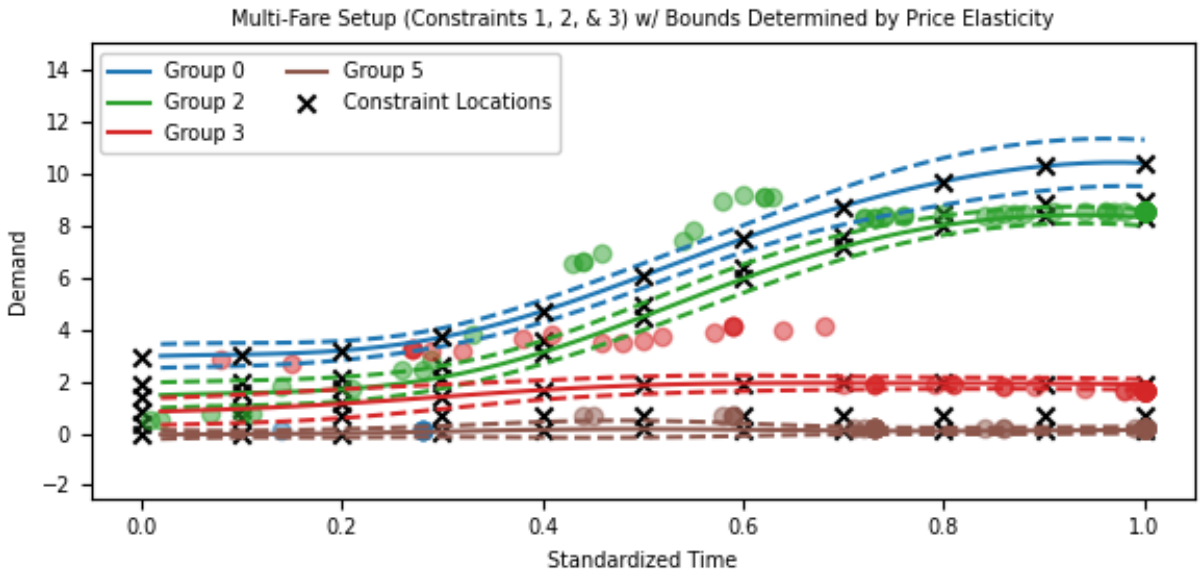


Figure 4.6: *CGP mean predictions and 2-standard deviation confidence intervals around the mean are shown for each group. The points shown represent each group's sample average of points calculated within 0.3 distance of the standardized time.*

4.8 Case Study: Airline Customer Buy-Down and Sell-Up Behavior

Here, a case study is presented to highlight how the CGPs described can also be used to model customer buy-down and sell-up behavior (described in Section 4.5). A multi-fare-multi-state CGP is trained to capture the buying behaviors of states E and C. In other words, the columns of $\boldsymbol{X}$ will represent both the fare group and state of the bookings observed. The first 5 columns will represent value of τ if $Z_j(m) = \mathrm{E}$ for all fare groups, while the remaining 5 columns will represent value of τ if $Z_j(m) = \mathrm{C}$. Each row will be designed to correspond when the observation is either $y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}})$ or $y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}})$ as to separate the demand arrival behaviors from one another. Each will use the parameters described in Section 4.7, where all 3 constraints will be applied to both $y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}})$ or $y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}})$ respectively, and Equation 4.12 will be used for the lower bounds of Constraint 2 using the same prices that Figure 4.5c assumed. One further constraint will be included describing the condition that $y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}}) \leq y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}})$, which will relay communication between the two states. This constraint will be useful as to ensure the equations defining the probability of buy-up and sell-down will also remain true.

The results of the trained CGP can be observed in Figure 4.7. In this figure, the top graph shows the data broken into states $y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}})$ and $y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}})$, along with the trained CGP predictions to the data. It is clear that the new constraint $y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}}) \leq y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}})$ is observable, as the predictions from $y_j^{\mathrm{C}}(\tau_j^{\mathrm{C}})$ are generally higher than $y_j^{\mathrm{E}}(\tau_j^{\mathrm{E}})$'s predictions. The bottom two graphs show the estimation of the probability of group buy-down and sell-up respectively versus allocated time. Group 5 is omitted from these probabilities as the sell-up probability could not be calculated for this group.

Interestingly, both probabilities tend to increase as time increases. This makes sense, as the chance of a customer buying down or selling up will naturally occur if time approaches infinity. However, the rates of each probability increasing differ. A 0.5 threshold is also plotted, which highlights how the probability of buy-down for Groups 3 and 4 is much more likely to occur in comparison any of the other buy-down and sell-up probabilities of other groups. Hence, this demonstrates how the outputs of the designed CGP is not only useful in predicting and relating different types of demand with one another, but also capable of estimating customer behavior between different groups. Again, this knowledge of probability of buy-down and sell-up is very useful to know when building optimization procedures to decide how much time should be allocated to each fare group, and the use of CGPs for estimating demand makes this possible.

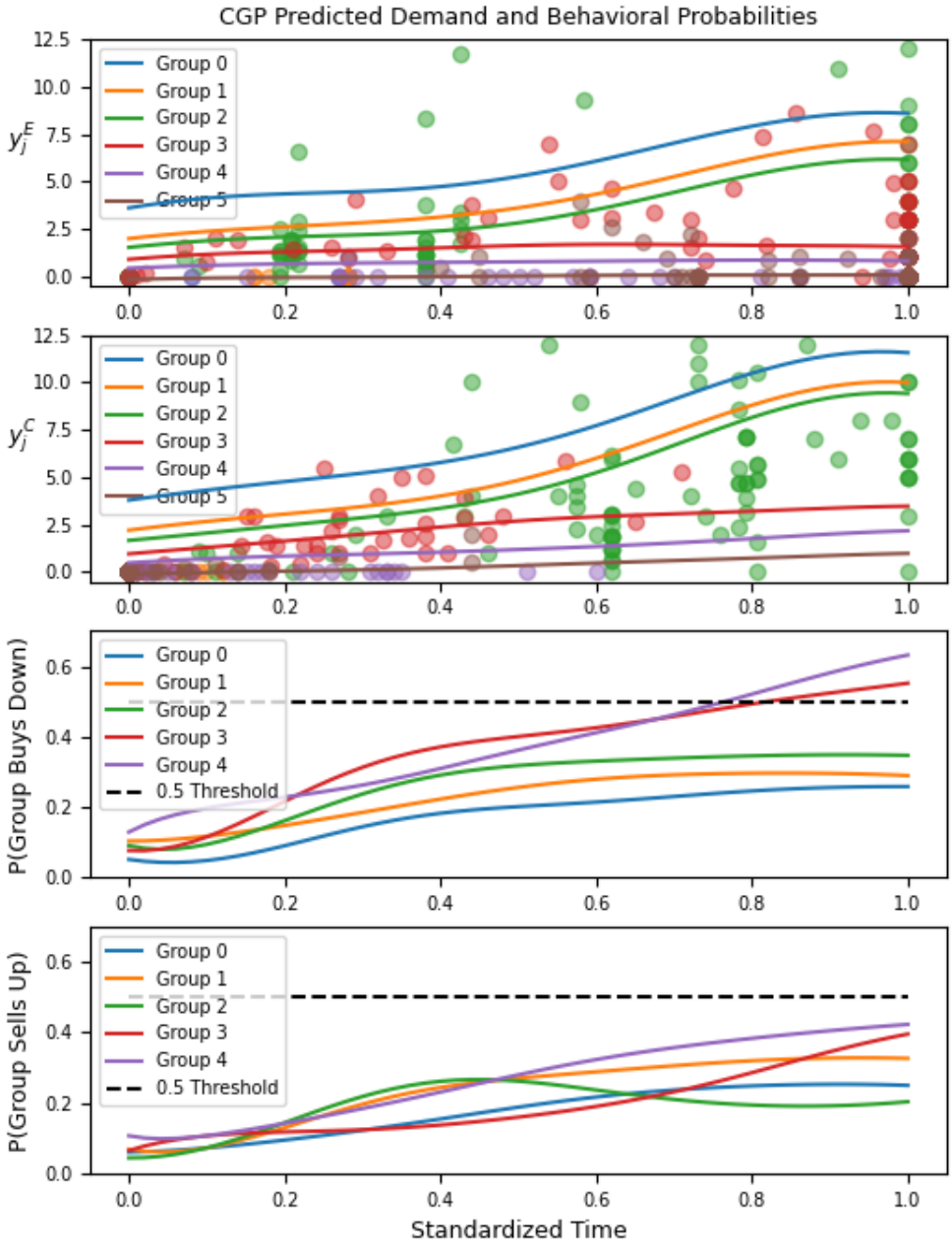


Figure 4.7: *The results of the multi-fare-multi-state constrained Gaussian Process. The top two graphs shows the CGP's predictions for data belonging to the E and C states respectively. The demand estimates are then converted into buy-down and sell-up probabilities via Equations 4.18 and 4.20 and shown in the bottom two graphs.*

4.9 Conclusion

This paper presents a demand setting where a constrained Gaussian Process can be beneficial for estimation. The demand setting involves estimating demand for different fare groups under the assumption of customer buy-down and sell-up behavior, which results in the observed data being partitioned into different states. If this partitioning is considered in conjunction with an airline's allocated time towards selling each fare group, this often results expensive fare groups having in little-to-no data being available for observations coming from the cheap state. Hence, estimation of demand under a single fare set-up is difficult to perform in practice. This paper avoids this problem by setting up the problem to consider multiple fare groups together for estimation. The setup is presented via a constrained Gaussian Process. By including constraints in a Gaussian Process, information both between and within different fare groups can be used together to estimate demand for groups with little-to-no data.

This paper also defines three different constraints that can be used in a constrained Gaussian Process via different linear operators and bounds. The constraints are tested using both simulated data and realistic airline booking data. In the simulated data example, the CGP is able to nearly capture the true demand curve of the simulated data. In realistic airline example, the CGP is able to follow the booking pattern of the test data using the presented constraints. As such, the constraints presented in this paper can be used to improve estimation of demand with little-to-no data when other groups with lots of data exist.

CHAPTER 5

Conclusions and Future Research

5.1 Summary of Contributions

In conclusion, this dissertation unites three significant contributions that collectively contribute to our understanding of human behavior analysis. By merging insights from the detection of distracted driving behavior, labeling of normal humanistic driving behaviors, and modeling of humanistic buying behavior, this dissertation strives to shed light on the complex nature of human behavior in diverse contexts. Through these endeavors, we hope to pave the way for further advancements in this field and foster the development of solutions that can positively impact society.

A summary of the major results and new contributions with regard to each chapter is provided below:

1. A fusion of three different state-space motion models is presented in Chapter 2. The fusion, called the Autonomous Multiple Model (AMM) algorithm, demonstrates how predictions errors from the models can be used to fuse the different models together. This likelihood of the prediction errors weights each model's predictions allowing for each model to influence the next predicted state. Chapter 2 also presents a control-chart-based decision strategy to monitor the prediction errors and label humanistic distraction status. The control charts combine a Cumulative Sum (CUSUM) control chart and a Exponentially Weighted Moving Average (EWMA) control chart together to provide multi-level assessment of the human distraction behavior.

2. An enhanced inference algorithm for the Hierarchical Dirichlet Process Hidden Semi-Markov Model (HDP-HSMM) is developed to counter the inconsistency issues regarding estimation of the true number of states. These issues are well known to occur from Dirichlet Processes in literature, and the enhanced algorithm allows the HDP-HSMM to be more robust to this issue. The enhanced algorithm presented in Chapter 3 adds a procedure into the inference algorithm that randomly merges redundant states together before each iteration of the inference procedure. Chapter 3 demonstrates how the procedure results in faster convergence of

emission and duration parameters and additionally results in estimating the true number of states belonging to the data. In the case study with naturalistic driving data, the presented algorithm is shown to tend to result in more meaningful and consistent emission means. This implies better interpretability from the modeling that enhances the understanding of humanistic driving behaviors.

3. The notation and setup of for modeling multiple fare groups' demand is provided in Chapter 4. Under this notation, three common demand constraints are mathematically defined such that they can incorporated into a constrained Gaussian Process Regression model to allow for each group to communicate with one another. Furthermore, the economics's concept of price elasticity is utilized to adjust the bounds of the constraints to improve demand estimation. This dissertation then demonstrates how the results of this modeling can be transformed into customer buy-down and sell-up probabilities that change with respect to time. This contributes to the modeling of humanistic buying behavior in that these probabilities are generally difficult estimate.

5.2 Future Research

While this dissertation makes strides in contributing towards different fields of human behavior analysis, research is still expected with regards to each work:

- For the context of the Autonomous Multiple Model algorithm, further development of defining the different level of distractions is still recommended. The algorithm presented in this dissertation assumes the level of distraction is correlated with the alarms generated by the EWMA and CUSUM control charts, and the number of consecutive windows in which the alarms occur. However, the alarms described in this dissertation are only used for binary decisions as opposed to a more descriptive distraction decision. With that in mind, the work presented in this paper can naturally be converted to further research, as the EWMA and CUSUM charts are effective in capturing different process disruptions. EWMA charts are effective in detecting smaller and gradual shifts, while CUSUM charts are more suitable for detecting larger and abrupt shifts. Hence these methodologies can potentially allow for users to be more descriptive of the true cause of alarms. Further development may help improve the false alarm and detection rates.

- For the context of the described robust HDP-HSMM, further research needs to be done regarding demonstrating the accuracy of the transition probability matrices. This dissertation focuses on demonstrating how the number of estimated states can accurately be inferred,

which translates into improved change point estimation, improved emission estimation, and improved duration estimation. However the transition probabilities are never fully discussed in this work. Further development in demonstrating the accuracy of the transition probabilities may open up possibilities to combine the outputs of the HDP-HSMM with the AMM discussed in Chapter 2. Lastly, a formal proof which demonstrates convergence in this dissertation enhanced block sampling procedure is still necessary for further research. This work demonstrates through a numerical study that speed of convergence improves greatly, but no theorems are presented which quantify this improvement.

- For the context of the airline demand modeling via a constrained Gaussian Process, further development of defining the constraint bounds is still open to research. This work demonstrates how the concept of price elasticity can be used to better define the demand estimates from one price group to another, however domain knowledge may help in estimating the other constraint bounds. Furthermore, the exploration of how to utilize constraints between different markets groups may be of interest for aiding estimation for markets with little historical information. Regarding the estimated of buy-down and sell-up probabilities, it is recommended to explore how these probabilities can change the way optimization is performed for airline demand allocation. These probabilities allow for optimization procedures to now have knowledge of how different decisions change consumer behavior, which can allow for airlines to make more strategic decisions regarding demand allocation.

APPENDIX A

A.1 Derivations of Q and R matrices in CV, CA and $CTRA$ models

In the Constant Velocity model, $x_t = \{L_{x,t}, L_{y,t}, V_{x,t}, V_{y,t}\}^T$ are the kinematic state variables where each dimension denotes longitudinal position, lane position, longitudinal velocity, and lateral velocity, respectively. By assuming the velocity is constant within each sampling time interval $\Delta t = 0.1$, the dynamic model is written as:

$$\begin{pmatrix} L_{x,t+1} \\ L_{y,t+1} \\ V_{x,t+1} \\ V_{y,t+1} \end{pmatrix} = \begin{pmatrix} L_{x,t} + V_{x,t}\Delta t \\ L_{y,t} + V_{y,t}\Delta t \\ V_{x,t} \\ V_{y,t} \end{pmatrix} + w.$$

Since the transition function f is linear, the above equation can be rewritten using matrix representation:

$$\begin{pmatrix} L_{x,t+1} \\ L_{y,t+1} \\ V_{x,t+1} \\ V_{y,t+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} L_{x,t} \\ L_{y,t} \\ V_{x,t} \\ V_{y,t} \end{pmatrix} + w.$$

It is often the case that measurements or estimations of lane position and longitudinal velocity are available. The relationship between these signals and the state variables are captured by the observation equation as:

$$y_{t+1} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} L_{x,t+1} \\ L_{y,t+1} \\ V_{x,t+1} \\ V_{y,t+1} \end{pmatrix} + u = \begin{pmatrix} V_{x,t+1} \\ L_{y,t+1} \end{pmatrix} + u.$$

To model the process noise covariance matrix Q, we need to compare the CV model with a more realistic model where the longitudinal and lateral accelerations $a_{x,t}$ and $a_{y,t}$ follow normal distribution with mean zero and unknown standard deviations, which is to assume $a_{x,t} \sim N(0, \sigma_{a,x}^2)$ and $a_{y,t} \sim N(0, \sigma_{a,y}^2)$. The differences between the two models are calculated as:

$$
\begin{aligned}
w &= \begin{pmatrix} L_{x,t} + V_{x,t}\Delta t + \frac{1}{2}a_{x,t}\Delta t^2 \\ L_{y,t} + V_{y,t}\Delta t + \frac{1}{2}a_{y,t}\Delta t^2 \\ V_{x,t} + a_{x,t}\Delta t \\ V_{y,t} + a_{y,t}\Delta t \end{pmatrix} - \begin{pmatrix} L_{x,t} + V_{x,t}\Delta t \\ L_{y,t} + V_{y,t}\Delta t \\ V_{x,t} \\ V_{y,t} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{2}a_{x,t}\Delta t^2 \\ \frac{1}{2}a_{y,t}\Delta t^2 \\ a_{x,t}\Delta t \\ a_{y,t}\Delta t \end{pmatrix}.
\end{aligned}
$$

We further assume $\sigma_{a,x}$ and $\sigma_{a,y}$ are independent, therefore the covariance matrix Q_{CV} is derived as:

$$
Q_{CV} = \begin{pmatrix} \frac{1}{4}\Delta t^4\sigma_{a,x}^2 & 0 & \frac{1}{2}\Delta t^3\sigma_{a,x}^2 & 0 \\ 0 & \frac{1}{4}\Delta t^4\sigma_{a,y}^2 & 0 & \frac{1}{2}\Delta t^3\sigma_{a,y}^2 \\ \frac{1}{2}\Delta t^3\sigma_{a,x}^2 & 0 & \Delta t^2\sigma_{a,x}^2 & 0 \\ 0 & \frac{1}{2}\Delta t^3\sigma_{a,y}^2 & 0 & \Delta t^2\sigma_{a,y}^2 \end{pmatrix}.
$$

For the Constant Acceleration model, the state variables are $x_t = \{L_{x,t}, L_{y,t}, V_{x,t}, V_{y,t}, a_{x,t}, a_{y,t}\}^T$. By assuming the acceleration is constant within each Δt, the dynamic model is:

$$
\begin{aligned}
\begin{pmatrix} L_{x,t+1} \\ L_{y,t+1} \\ V_{x,t+1} \\ V_{y,t+1} \\ a_{x,t+1} \\ a_{y,t+1} \end{pmatrix} &= \begin{pmatrix} L_{x,t} + V_{x,t}\Delta t + \frac{1}{2}a_{x,t}\Delta t^2 \\ L_{y,t} + V_{y,t}\Delta t + \frac{1}{2}a_{y,t}\Delta t^2 \\ V_{x,t} + a_{x,t}\Delta t \\ V_{y,t} + a_{y,t}\Delta t \\ a_{x,t} \\ a_{y,t} \end{pmatrix} + w \\
&= \begin{pmatrix} 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 \\ 0 & 0 & 1 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} L_{x,t} \\ L_{y,t} \\ V_{x,t} \\ V_{y,t} \\ a_{x,t} \\ a_{y,t} \end{pmatrix} + w
\end{aligned}
$$

Similarly, we have the observation equation:

$$y_{t+1} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} L_{x,t+1} \\ L_{y,t+1} \\ V_{x,t+1} \\ V_{y,t+1} \\ a_{x,t+1} \\ a_{y,t+1} \end{pmatrix} + u = \begin{pmatrix} V_{x,t+1} \\ L_{y,t+1} \end{pmatrix} + u.$$

Instead of the constant acceleration assumption, we still assume the accelerations follow normal distribution with unknown standard deviations. The process noise covariance matrix is calculated similarly as:

$$Q_{CA} = \begin{pmatrix} \frac{1}{4}\Delta t^4 \sigma_{a,x}^2 & 0 & \frac{1}{2}\Delta t^3 \sigma_{a,x}^2 & 0 & \frac{1}{2}\Delta t \sigma_{a,x}^3 & 0 \\ 0 & \frac{1}{4}\Delta t^4 \sigma_{a,y}^2 & 0 & \frac{1}{2}\Delta t^3 \sigma_{a,y}^2 & 0 & \frac{1}{2}\Delta t \sigma_{a,y}^3 \\ \frac{1}{2}\Delta t^3 \sigma_{a,x}^2 & 0 & \Delta t^2 \sigma_{a,x}^2 & 0 & \Delta t \sigma_{a,x}^2 & 0 \\ 0 & \frac{1}{2}\Delta t^3 \sigma_{a,y}^2 & 0 & \Delta t^2 \sigma_{a,y}^2 & 0 & \Delta t \sigma_{a,y}^2 \\ \frac{1}{2}\Delta t \sigma_{a,x}^3 & 0 & \Delta t \sigma_{a,x}^2 & 0 & \sigma_{a,x}^2 & 0 \\ 0 & \frac{1}{2}\Delta t \sigma_{a,y}^3 & 0 & \Delta t \sigma_{a,y}^2 & 0 & \sigma_{a,y}^2 \end{pmatrix}$$

In the CTRA model, the state variables are $x_t = \{L_{x,t}, L_{y,t}, \theta_t, V_t, \omega_t, a_t\}^T$ where θ_t denotes vehicle heading, V_t denotes velocity, ω_t denotes yaw rate and a_t denotes acceleration. Within each time interval, the acceleration and yaw rate are assumed to be constant. The transition equation is written as:

$$\begin{pmatrix} L_{x,t+1} \\ L_{y,t+1} \\ \theta_{t+1} \\ V_{t+1} \\ \omega_{t+1} \\ a_{t+1} \end{pmatrix} = f\left(\begin{pmatrix} L_{x,t} \\ L_{y,t} \\ \theta_t \\ V_T \\ \omega_t \\ a_t \end{pmatrix}\right)$$

$$= \begin{pmatrix} L_{x,t} + \frac{V_t}{\omega_t}\left(sin(\theta_t + \omega_t \Delta t) - sin\theta_t\right) + \frac{a_t \Delta t}{\omega_t} sin\left(\theta_t + \omega_t \Delta t\right) - \frac{a_t}{\omega_t^2} cos\theta_t + \frac{a_t}{\omega_t^2} cos(\theta_t + \omega_t \Delta t) \\ L_{y,t} + \frac{V_t}{\omega_t}\left(cos\theta_t - cos(\theta_t + \omega_t \Delta t)\right) - \frac{a \Delta t}{\omega_t} cos(\theta_t + \omega_t \Delta t) + \frac{a_t}{\omega_t^2} sin(\theta_t + \omega_t \Delta t) - \frac{a_t}{\omega_t^2} sin\theta_t \\ \theta_t + \omega_t \Delta t \\ V_t + a_t \Delta t \\ \omega_t \\ a_t \end{pmatrix}.$$

The Extended Kalman filter approximation is used since the transition function is not linear. The approximation matrix is calculated as the Jacobian matrix of transition function f, which is:

$$F = \begin{pmatrix} 1 & 0 & a_1 & a_2 & a_3 & a_4 \\ 0 & 1 & a_5 & a_6 & a_7 & a_8 \\ 0 & 0 & 1 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where:

$$\begin{aligned}
a_1 &= \frac{V_t}{\omega_t}\left(cos(\theta_t + \omega_t\Delta t) - cos\theta_t\right) + \frac{a_t\Delta t}{\omega_t}cos(\theta_t + \omega\Delta t) + \frac{a_t}{\omega_t^2}\left(sin\theta_t - sin(\theta_t + \omega\Delta t)\right), \\
a_2 &= \frac{1}{\omega_t}\left(sin(\theta_t + \omega_t\Delta t) - sin\theta_t\right), \\
a_3 &= \frac{1}{\omega_t^2}\left[V_t sin\theta_t + V_t cos(\theta_t + \omega_t\Delta t)\omega_t\Delta t - V_t sin(\theta_t + \omega_t\Delta_t)\right] \\
&\quad - \frac{1}{\omega_t^2}\left[a_t\Delta t^2 cos(\theta_t + \omega_t\Delta t) - a_t\Delta t sin(\theta_t + \omega_t\Delta t)\right] + \frac{2acos\theta_t}{\omega_t^3} \\
&\quad - \frac{1}{\omega^4}\left[a_t sin(\theta_t + \omega_t\Delta t)\Delta t\omega_t^2 + 2a_t cos(\theta_t + \omega_t\Delta t)\omega_t\right], \\
a_4 &= \frac{\Delta t}{\omega_t}sin(\theta_t + \omega_t\Delta t) - \frac{1}{\omega_t^2}cos\theta_t + \frac{1}{\omega_t^2}cos(\theta_t + \omega_t\Delta t), \\
a_5 &= \frac{V_t}{\omega_t}\left(sin(\theta_t + \omega_t\Delta t) - sin\theta_t\right) + \frac{a_t\Delta t}{\omega_t}sin(\theta_t + \omega_t\Delta t) + \frac{a_t}{\omega_t^2}\left[cos(\theta_t + \omega_t\Delta t) - cos\theta_t\right], \\
a_6 &= \frac{1}{\omega_t}\left(cos\theta_t - cos(\theta_t + \omega_t\Delta t)\right), \\
a_7 &= \frac{1}{\omega_t^2}\left[V_t cos(\theta_t + \omega_t\Delta t) - V_t cos\theta_t - V_t sin(\theta_t + \omega_t\Delta t)\omega_t\Delta t\right] \\
&\quad + \frac{1}{\omega_t^2}\left[a_t\Delta t^2 sin(\theta_t + \omega_t\Delta t) + a\Delta t cos(\theta_t + \omega_t\Delta t)\right] + \frac{2a_t sin\theta_t}{\omega_t^3} \\
&\quad + \frac{1}{\omega_t^4}\left[a_t cos(\theta_t + \omega_t\Delta t)\Delta t\omega_t^2 - 2a_t sin(\theta_t + \omega_t\Delta t)\omega_t\right], \\
a_8 &= -\frac{\Delta t}{\omega_t}cos(\theta_t + \omega_t\Delta t) + \frac{1}{\omega_t^2}\left[sin(\theta_t + \omega_t\Delta t) - sin\theta_t\right].
\end{aligned}$$

Lane position, velocity and yaw rate are measured or estimated in the application, therefore:

$$y_{t+1} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} L_{x,t+1} \\ L_{y,t+1} \\ \theta_{t+1} \\ V_{t+1} \\ \omega_{t+1} \\ a_{t+1} \end{pmatrix} + u = \begin{pmatrix} L_{y,t+1} \\ V_{t+1} \\ \omega_{t+1} \end{pmatrix} + u$$

To calculate the process noise matrix, we assume the actual acceleration and yaw rate follow normal distribution with unknown standard deviation σ_a^2 and σ_ω^2. The process noise matrix is written as:

$$Q_{CTRA} = A \begin{pmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_\omega^2 \end{pmatrix} A^T,$$

where:

$$A = \begin{pmatrix} a_4 & a_3 \\ a_8 & a_7 \\ 0 & \Delta t \\ \Delta t & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

A.2 Updating Algorithm of Kalman Filter

Let $\hat{x}_{t+1|t}$ and $P_{t+1|t}$ denote the mean and covariance of prediction (pre-fit mean and covariance) at time $t+1$ given all the historical data until t. When the data at time $t+1$ comes, the updated mean and covariance (post-fit mean and covariance) is denoted by $\hat{x}_{t+1|t+1}$ and $P_{t+1|t+1}$. Suppose the transition function is linear with matrix F and the observation matrix is G. The Kalman filter is updated as follows.

In the prediction step, the pre-fit mean and covariance at time $t+1$ are generated based on the post-fit mean and covariance at the previous time point t as:

$$\hat{x}_{t+1|t} = F\hat{x}_{t|t},$$

$$P_{t+1|t} = FP_{t|t}F^T + Q.$$

By comparing the prediction and observation, we obtain the residual and its pre-fit covariance:

$$\epsilon_{t+1|t} = y_{t+1} - G\hat{x}_{t+1|t},$$

$$S_{t+1} = R + GP_{t|t}G^T.$$

Then the information of time $t + 1$ is used to update the Kalman filter. The updated term is known as Kalman gain, which is calculated as:

$$K_{t+1} = P_{t+1|t}G^T S_{t+1}^{-1}.$$

The Kalman gain is then used to update the post-fit mean and covariance at time $t + 1$. They are:

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + K_{t+1}\epsilon_{t+1|t},$$

$$P_{t+1|t+1} = (I - K_{t+1}G)P_{t+1|t}.$$

The procedure is repeated iteratively till the end of the data stream.

BIBLIOGRAPHY

[1] Lars Aberg and Per-Arne Rimmo. Dimensions of aberrant driver behaviour. *Ergonomics*, 41(1):39–56, 1998.

[2] Christian Agrell. Gaussian processes with linear operator inequality constraints. *arXiv preprint arXiv:1901.03134*, 2019.

[3] Matthew J Beal, Zoubin Ghahramani, and Carl E Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2002.

[4] Luis M Bergasa, Daniel Almería, Javier Almazán, J Javier Yebes, and Roberto Arroyo. Drivesafe: An app for alerting inattentive drivers and scoring driving behaviors. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 240–245. IEEE, 2014.

[5] E Andrew Boyd and Royce Kallesen. Practice papers: The science of revenue management when passengers purchase the lowest available fare. *Journal of Revenue and Pricing Management*, 3(2):171–177, 2004.

[6] Christopher Andrew Boyer. *Statistical methods for forecasting and estimating passenger willingness-to-pay in airline revenue management*. PhD thesis, Massachusetts Institute of Technology, 2010.

[7] Stefano Di Cairano, Daniele Bernardini, Alberto Bemporad, and Ilya V Kolmanovsky. Stochastic mpc with learning for driver-predictive vehicle control and its application to hev energy management. *IEEE Transactions on Control Systems Technology*, 22(3):1018–1031, 2013.

[8] Thomas A Dingus, Sheila G Klauer, Vicki L Neale, Andy Petersen, Suzanne E Lee, JD Sudweeks, Miguel A Perez, Jonathan Hankey, DJ Ramsey, Santosh Gupta, et al. The 100-car naturalistic driving study, phase ii-results of the 100-car field experiment. Technical report, 2006.

[9] Alberto Fernández, Rubén Usamentiaga, Juan Luis Carús, and Rubén Casado. Driver distraction using visual-based sensors and algorithms. *Sensors*, 16(11):1805, 2016.

[10] Thomas Fiig, Larry R Weatherford, and Michael D Wittman. Can demand forecast accuracy be linked to airline revenue? *Journal of Revenue and Pricing Management*, 18(4):291–305, 2019.

[11] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. An hdp-hmm for systems with state persistence. In *Proceedings of the 25th international conference on Machine learning*, pages 312–319. ACM, 2008.

[12] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, pages 1020–1056, 2011.

[13] Gregory Fridman and Maria Lapina. Maximum likelihood approach for demand unconstraining problem with censoring information incompleteness. *Journal of Revenue and Pricing Management*, 15(1):37–51, 2016.

[14] G Gallego, L Li, and R Ratliff. Revenue management with customer choice models. Technical report, Technical report, Presentation at AGIFORS RMD and Cargo Study Group, South Korea, 2007.

[15] Elisabeth Gassiat, Judith Rousseau, et al. About the posterior distribution in hidden markov models with unknown number of states. *Bernoulli*, 20(4):2039–2075, 2014.

[16] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[17] Andrew Gelman and Donald B Rubin. A single series from the gibbs sampler provides a false sense of security. *Bayesian statistics*, 4:625–631, 1992.

[18] Aritra Guha, Nhat Ho, and XuanLong Nguyen. On posterior contraction of parameters and interpretability in bayesian mixture modeling. *arXiv preprint arXiv:1901.05078*, 2019.

[19] Peng Guo, Baichun Xiao, and Jun Li. Unconstraining methods in revenue management systems: Research overview and prospects. *Advances in Operations Research*, 2012, 2012.

[20] Douglas M Hawkins. A cusum for a scale parameter. *Journal of Quality Technology*, 13(4):228–231, 1981.

[21] C Hopperstad and P Belobaba. Alternative rm algorithms for unrestricted fare structures. In *AGIFORS Reservation and Yield Management Meeting. Aukland, New Zealand*, pages 28–31, 2004.

[22] Ajay Jasra, Chris C Holmes, and David A Stephens. Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67, 2005.

[23] Matthew J Johnson and Alan S Willsky. Bayesian nonparametric hidden semi-markov models. *Journal of Machine Learning Research*, 14(Feb):673–701, 2013.

[24] Matthew James Johnson and Alan S Willsky. Dirichlet posterior sampling with truncated multinomial likelihoods. *arXiv preprint arXiv:1208.6537*, 2012.

[25] Michael I Jordan. Bayesian nonparametric learning: Expressive priors for intelligent systems. *Heuristics, probability and causality: A tribute to Judea Pearl*, 11:167–185, 2010.

[26] Michael I Jordan and Yee Whye Teh. A gentle introduction to the dirichlet process, the beta process, and bayesian nonparametrics. *Dept. Statistics, UC Berkeley*, 2014.

[27] Simon J Julier and Jeffrey K Uhlmann. A general method for approximating nonlinear transformations of probability distributions. Technical report, Technical report, Robotics Research Group, Department of Engineering Science, University of Oxford, 1996.

[28] Simon J Julier and Jeffrey K Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[29] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45, 1960.

[30] Alex Paul Kamson, LN Sharma, and S Dandapat. Multi-centroid diastolic duration distribution based hsmm for heart sound segmentation. *Biomedical signal processing and control*, 48:265–272, 2019.

[31] Sandeep Karmarkar, Dutta Goutam, and Bandyopadhyay Tathagata. Revenue impacts of demand unconstraining and accounting for dependency. *Journal of Revenue and Pricing Management*, 10(4):367–381, 2011.

[32] Hyungsul Kim, Manish Marwah, Martin Arlitt, Geoff Lyon, and Jiawei Han. Unsupervised disaggregation of low frequency power measurements. In *Proceedings of the 2011 SIAM international conference on data mining*, pages 747–758. SIAM, 2011.

[33] J Zico Kolter and Matthew J Johnson. Redd: A public data set for energy disaggregation research. In *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA*, volume 25, pages 59–62, 2011.

[34] D Magill. Optimal adaptive estimation of sampled stochastic processes. *IEEE Transactions on Automatic Control*, 10(4):434–439, 1965.

[35] Clara Marina Martinez, Mira Heucke, Fei-Yue Wang, Bo Gao, and Dongpu Cao. Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):666–676, 2017.

[36] Jeffrey W Miller and Matthew T Harrison. A simple example of dirichlet process mixture inconsistency for the number of components. *arXiv preprint arXiv:1301.2708*, 2013.

[37] Douglas C Montgomery. *Introduction to statistical quality control*. John Wiley & Sons, 2007.

[38] Kevin P Murphy. Hidden semi-markov models (hsmms). *unpublished notes*, 2, 2002.

[39] Emily Nodine, Andy Lam, Scott Stevens, Michael Razo, Wassim Najm, et al. Integrated vehicle-based safety systems (ivbss) light vehicle field operational test independent evaluation. Technical report, United States. National Highway Traffic Safety Administration, 2011.

[40] World Health Organization et al. Global status report on road safety 2013: supporting a decade of action: summary. 2013.

[41] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[42] Athanasios Papoulis and S Unnikrishna Pillai. *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[43] Ryan R Pitre, Vesselin P Jilkov, and X Rong Li. A comparative study of multiple-model algorithms for maneuvering target tracking. In *Signal Processing, Sensor Fusion, and Target Recognition XIV*, volume 5809, pages 549–561. International Society for Optics and Photonics, 2005.

[44] Ilan Price, Jaroslav Fowkes, and Daniel Hopman. Gaussian processes for demand unconstraining. *arXiv preprint arXiv:1711.10910*, 2017.

[45] Lawrence Rabiner and B Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[46] Mizanur Rahman, Mashrur Chowdhury, Kakan Dey, M Rafiul Islam, and Taufiquar Khan. Evaluation of driver car-following behavior models for cooperative adaptive cruise control systems. *Transportation Research Record*, 2622(1):84–95, 2017.

[47] CE Rasmussen. Rasmussen and williams: Gaussian processes for machine learning. *J. Mach. Learn. Res.*, 11:3011–3015, 2006.

[48] Branko Ristic, Sanjeev Arulampalam, and Neil Gordon. *Beyond the Kalman filter: Particle filters for tracking applications*. Artech house, 2003.

[49] Fridulv Sagberg, Selpi, Giulio Francesco Bianchi Piccinini, and Johan Engström. A review of research on driving styles and road safety. *Human factors*, 57(7):1248–1275, 2015.

[50] Robin Schubert, Eric Richter, and Gerd Wanielik. Comparison and evaluation of advanced motion models for vehicle tracking. In *Information Fusion, 2008 11th International Conference on*, pages 1–6. IEEE, 2008.

[51] Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

[52] Shadi Sharif Azadeh, Patrice Marcotte, and Gilles Savard. A taxonomy of demand uncensoring methods in revenue management. *Journal of Revenue and Pricing Management*, 13(6):440–456, 2014.

[53] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2017.

[54] Matthew Sperrin, Thomas Jaki, and Ernst Wit. Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing*, 20(3):357–366, 2010.

[55] Catalina Stefanescu. Multivariate customer demand: modeling and estimation from censored sales. *Available at SSRN 1334353*, 2009.

[56] Rui Sun. *An integrated solution based irregular driving detection*. Springer, 2016.

[57] Rui Sun, Washington Yotto Ochieng, and Shaojun Feng. An integrated solution for lane level irregular driving detection on highways. *Transportation Research Part C: Emerging Technologies*, 56:61–79, 2015.

[58] Laura Symul and Susan P Holmes. Labeling self-tracked menstrual health records with hidden semi-markov models. *medRxiv*, 2021.

[59] Yee Whye Teh. Dirichlet process., 2010.

[60] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[61] John R Treat. A study of precrash factors involved in traffic accidents. *HSRI Research Review*, 1980.

[62] Wenshuo Wang, Junqiang Xi, and Ding Zhao. Driving style analysis using primitive driving patterns with bayesian nonparametric approaches. *IEEE Transactions on Intelligent Transportation Systems*, 20(8):2986–2998, 2018.

[63] Wenshuo Wang and Ding Zhao. Extracting traffic primitives directly from naturalistically logged data for self-driving applications. *IEEE Robotics and Automation Letters*, 3(2):1223–1229, 2018.

[64] Larry Weatherford. The history of unconstraining models in revenue management. *Journal of Revenue and Pricing Management*, 15(3):222–228, 2016.

[65] Larry R Weatherford and Stefan Pölt. Better unconstraining of airline demand data in revenue management systems for improved forecast accuracy and greater revenues. *Journal of Revenue and Pricing Management*, 1(3):234–254, 2002.

[66] Larry R Weatherford and Richard M Ratliff. Review of revenue management methods with dependent demands. *Journal of Revenue and Pricing Management*, 9(4):326–340, 2010.

[67] Chuck Wooters and Marijn Huijbregts. The icsi rt07s speaker diarization system. In *Multimodal Technologies for Perception of Humans*, pages 509–519. Springer, 2007.

[68] Richard Henry Zeni. *Improved forecast accuracy in revenue management by unconstraining demand estimates from censored data*. Rutgers The State University of New Jersey-Newark, 2001.

[69] Ding Zhao, Yaohui Guo, and Yunhan Jack Jia. Trafficnet: An open naturalistic driving scenario library. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2017.